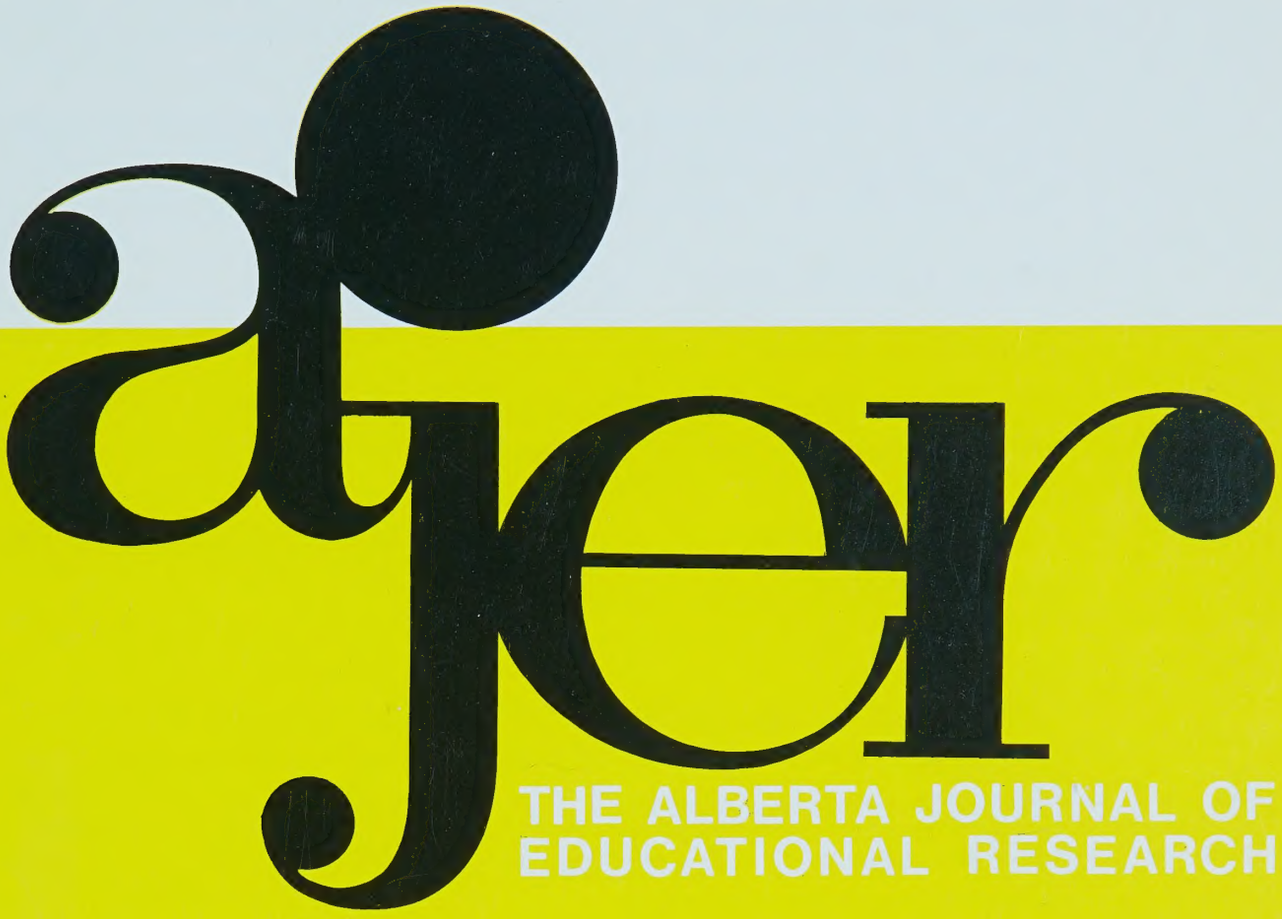

THE ALBERTA JOURNAL OF EDUCATIONAL RESEARCH

VOLUME XXXVIII NUMBER 3 SEPTEMBER 1992

PUBLISHED BY
THE UNIVERSITY OF ALBERTA • EDMONTON

The Alberta Journal of Educational Research

Published in March, June, September, and December
by the Faculty of Education, University of Alberta

A quarterly journal devoted to the dissemination, criticism, interpretation, and encouragement of all forms of systematic enquiry into education and fields related to or associated with education.

Editor: Robert H. Short

Consulting Editors

James Chapman
Massey University

Rod Clifton
University of Manitoba

Dianne Common
Lakehead University

Carol Crealock
University of Western Ontario

John Hattie
University of Western Australia

Billie Housego
University of British Columbia

Linda Larocque
University of Alberta

Stewart J.H. McCann
University College of Cape Breton

Les McLean
Ontario Institute for Studies in Education

Linda Phillips
Memorial University of Newfoundland

Sharon Robertson
University of Calgary

Faculty Advisory Committee

John G. Paterson (Chair)
Fern Snart
Joyce Edwards
Graham Fishburne

AJER gratefully acknowledges support from the Social Sciences and Humanities Research Council of Canada, from the Alberta Advisory Committee for Educational Studies, and from the Office of Research Services, University of Alberta.

The subscription rate is $32.00 per year for individuals, $40.00 per year for institutions. Add $8.00 for delivery outside Canada. Single copies are $8.00 each. Subscriptions and sales in Canada will be charged 7% GST. Please make cheques payable to *The Alberta Journal of Educational Research*. Back issues are available; rates supplied on request. Claims for undelivered copies must be received within three months of publication.

Address all communications and manuscript submissions to *The Alberta Journal of Educational Research*, Faculty of Education, Publication Services, 4-116 Education North, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5.

Publications Mail Registration Number 1436

Faculty of Education
University of Alberta

The Alberta Journal of Educational Research

Volume XXXVIII, Number 3, September 1992

Stephen P. Norris	**155**	A Demonstration of the Use of Verbal Reports of Thinking in Multiple-Choice Critical Thinking Test Design
Judith K. Bernhard	**177**	Gender-Related Attitudes and the Development of Computer Skills: A Preschool Intervention
Ellen M. Regan *John A. Ross*	**189**	Professional Development for School Consultants: An Evaluation
Marziya Yasmin *Baha Abu-Laban*	**205**	Ethnicity and Occupational Inequality: A Reconsideration
Estelle S. Gellman *Mina Berkowitz*	**219**	Factors Perceived as Important in Teacher Evaluation
Dushyanta Persaud *Paul R. Madak*	**235**	Graduates and Dropouts: Comparing Perceptions of Self, Family, and School Supports

ISSN 0002-4805

The Alberta Journal of Educational Research *Vol. XXXVIII, No. 3, September 1992, 155-176*

Stephen P. Norris
Memorial University of Newfoundland

A Demonstration of the Use of Verbal Reports of Thinking in Multiple-Choice Critical Thinking Test Design

A methodology is described for using verbal reports of thinking to develop and validate multiple-choice tests of critical thinking. The procedure includes the following steps: devising normative models of thinking for each item; developing a method for interviewing subjects; collecting verbal reports of thinking from samples of subjects; using the normative models to rate the quality of the thinking portrayed in the reports; comparing the quality of thinking to answer choices and accumulating those comparisons across subjects for each item; identifying and modifying suspect items. The methodology is offered as a way to gather evidence on the truth of claims that currently available multiple-choice critical thinking tests do not measure critical thinking (McPeck, 1981; Petrie, 1986; Whimbey, 1985), and as a way to develop new critical thinking tests.

On présente ici une méthodologie d'utilisation des rapports verbaux sur la pensée pour développer et valider des tests à choix multiples de pensée critique. Le procédé comprend les étapes suivantes: conception de modèles normatifs de pensée pour chaque item; développement d'une méthode d'entrevue avec les sujets; collecte de rapports verbaux sur la pensée utilisant des échantillons de sujets; utilisation des modèles normatifs pour évaluer la qualité de la pensée telle qu'elle ressort des rapports; comparaison de la qualité de la pensée aux choix de réponses, et accumulation de ces comparaisons pour tous les sujets pour chaque item; identification et modification des items suspects. La méthodologie est offerte comme un moyen d'obtenir de l'évidence sur la valeur des idées selon lesquelles les tests à choix multiples actuels ne mesurent pas la pensée critique (McPeck, 1981; Petrie, 1986; Whimbey, 1985), et comme un moyen de développer de nouveaux tests de pensée critique.

During the last decade interest has grown in teaching and assessing critical thinking (Ennis & Norris, 1990). However, few assessment instruments are available. Furthermore, most of those available are multiple choice, and many advocates of critical thinking are dissatisfied in principle with multiple-choice testing formats (e.g., McPeck, 1981; Petrie, 1986; Whimbey, 1985). Such advocates insist that critical thinking tests must provide direct evidence on the thinking processes of examinees who take the tests, and multiple-choice critical thinking tests typically do not give such evidence. However, multiple-choice tests are popular because of their efficiency (Ennis & Norris, 1990) and are suited particularly well to some assessment tasks (Norris, 1989). Therefore, if multiple-choice testing of critical thinking could be justified (and seen as justified, because one cannot ignore perceptions) through design procedures that provided more direct evidence on the thinking processes that the tests

---

Stephen Norris is a professor of educational research and philosophy of education. In addition to critical thinking testing, his interests include the theory of reading comprehension, and philosophy of science and science education.

elicit, then both the critics and some important assessment goals might be satisfied.

This article describes how verbal reports of examinees' thinking on draft items of multiple-choice critical thinking tests can be used systematically to provide such direct validity evidence. Direct evidence on thinking processes gathered at the test design stage can be used to support more strongly inferences about examinees' thinking processes made at the use stage than design procedures that do not gather such direct evidence. Because direct evidence on thinking processes can play such an important role in test interpretation, testing theorists have endorsed verbal reports of thinking for amplifying the meaning of the constructs a test is measuring (Cronbach, 1971), for representing those constructs by the mental processes that underlie performance (Embretson, 1983; Embretson, Schneider, & Roth, 1986; Haney & Scott, 1987), for directly analyzing the processes underlying item performance (Messick, 1989), and for specifying the intellectual processes used to perform test tasks (Anastasi, 1988). Such endorsements support the premise of this article, namely, that the use of verbal reports of thinking can help meet the criticism that there is no direct evidence that multiple-choice critical thinking tests in fact measure critical thinking.

Despite the theoretical support, there are very few reported uses of verbal reports of thinking in test development and validation. When verbal reports have been used, the information in them rarely has been analyzed and compiled systematically. Rather, as will be shown in the section on past use of verbal reports of thinking in test design, the information usually has been used casually to support intuitive judgments about whether items mislead, contain unwanted clues, use unknown vocabulary, and the like. These are important considerations in test design, but it would be useful to see how verbal reports of thinking might be used systematically to help make construct validity judgments about multiple-choice critical thinking tests.

This article reports how verbal reports of thinking were used to develop and validate a multiple-choice critical thinking test of judging the credibility of observations. With suitable adjustments, the approach could be used to develop and to examine the validity of multiple-choice tests of other aspects of critical thinking.

Background

The Role of Observation in Critical Thinking

For the purposes of this study, Ennis's (1987) definition of critical thinking was adopted: "Critical thinking is reasonable reflective thinking that is focused on deciding what to believe or do" (p. 10). According to Ennis, thinking critically starts from a basis of information (information from others, from observation, and from conclusions previously drawn), and, through the process of inference, moves from that basis to decisions about belief and action. Observations are central to critical thinking on this model, because they are part of the basis on which decisions are based.

Motivation

Most definitions of critical thinking, including Ennis's, focus more on the process of thinking than its products. Paul (1985) argues that for learning to be

rational "the process that leads to belief is more important than belief itself" (p. 5) It is a result of this focus on process that many critical thinking theorists doubt that traditional multiple-choice tests can measure critical thinking, since such tests reveal products of thinking, not thinking itself.

This doubt is not pedantic. Criticisms of existing multiple-choice critical thinking tests have shown that keyed answers can be reached by thinking uncritically and unkeyed answers by thinking critically (Ennis & Norris, 1990; Norris, 1991). Such criticisms are usually hypothetical and not supported by empirical evidence on how examinees choose their answers. The point of the criticisms, however, is that there is no direct evidence one way or the other. Given the lack of evidence, there is a perceived unacceptable risk that some examinees are rewarded unfairly and others penalized unfairly. Such perceptions lead to a skepticism about the suitability of multiple-choice tests for measuring critical thinking.

A similar conclusion about multiple-choice critical thinking tests follows from research on thinking in complex domains and in ill-structured problems. Such thinking can follow multiple paths to multiple solutions (Resnick, 1987; Spiro, Vispoel, Schmitz, Samarapungavan, & Boerger, 1987), and often there is no expert consensus on either a correct path or a correct solution. Thus one might wonder how multiple-choice tests, which typically demand one solution, can test for critical thinking in complex domains.

Advocates of critical thinking are concerned, therefore, that multiple-choice testing cannot serve well the goal of critical thinking evaluation. On the other hand, multiple-choice tests seem an appropriate format for testing certain aspects of critical thinking, such as using the many principles for appraising the credibility of information, and there is considerable interest in students' ability to appraise information as opposed to merely locating it (e.g., National Assessment of Educational Progress, 1985). So it is important to know whether multiple-choice critical thinking tests can be designed to provide more direct evidence on the thinking that underlies performance. Such direct evidence on construct validity would place these tests on a sounder theoretical and empirical footing. Eliciting verbal reports of thinking on trial test items as described subsequently is one plausible way to collect the direct evidence desired.

Despite the references made previously to direct evidence, construct validity must be inferred (Cronbach, 1971). However, inferences about thinking processes based on item statistics and content analysis are *more* indirect and hence more untrustworthy than is desirable (Connolly & Wantman, 1964). Schuman (1966), for instance, has demonstrated that item performance can show excellent variation, correlate well with other item performances, and be related to other variables in relevant ways, even though the majority of examinees do not understand the item. As a result, some testing specialists recommend seeking more direct evidence on examinees' thinking than that contained in item statistics, such as evidence from examinees' verbal reports of their thinking as they work through test items (Anastasi, 1988; Haney & Scott, 1987) and from an "analysis of reasons" (Messick, 1989). Evidence of this sort can support inferences rarely supportable by item analysis data alone.

Eliciting verbal reports of examinees' thinking is an attempt to gain information on the mental processes leading to their performance on the test, that is, to learn the information, mental strategies, and reasoning principles they use in choosing answers. The aim is to look beneath the surface of answer choices in order to gain more direct insight into the reasons for the choices, so that more trustworthy inferences can be made about the causes of performance and, derivatively, about construct validity. In the specific context of testing for critical thinking, the aim would be to use information in examinees' verbal reports of thinking to judge whether it is differences in critical thinking, and not something else, that account for differences in scores. In order for the procedure to work, a way of eliciting verbal reports and of using the information contained in them needs to be designed and justified.

Past Use of Verbal Reports of Thinking in Test Design

Verbal reports of thinking have not been used often in test design. When such studies have been reported, direction on how the information in verbal reports might be used to quantify the quality of examinees' thinking, on how to make comparisons between the quality of thinking as displayed in examinees' verbal reports and their choices of answers, and on how to accumulate and coordinate these comparisons across several examinees to help judge the validity of tests has been inadequate.

Bloom and Broder (1950) were among the earliest users of verbal reports of thinking in test development. They asked a sample of college students to think aloud while they worked through questions on a problem solving test. The verbal reports yielded "relatively consistent and meaningful data from the majority of subjects" (p. 90) and helped to support inferences from the students' problem solutions to the thinking that led to the solutions. However, attempts were eschewed explicitly to formalize the procedures for gathering and using the verbal report information.

Kropp (1956) used verbal reports of thinking to examine the relationship between choosing correct answers on 10 multiple-choice items and the reasoning used to make the choices. The items were designed to measure understanding of text with physical science content and were judged to involve more complex thinking than simply recalling information or locating it in the text. The verbal reports showed that a great variety of reasoning processes led to choosing correct answers. Because some of these processes were not justifiable, such as outright guessing, Kropp concluded that it is hazardous to infer the quality of reasoning processes on the basis of answers chosen. However, no procedure was described for determining the *tendency* for unjustified reasoning to lead to correct answers across the sample of examinees for each item.

McGuire (1963) used verbal reports of thinking to construct medical school certification examinations that test for critical thinking and problem solving and not merely factual recall. Experts in the field were asked to classify test questions according to a modified Bloom's taxonomy. Medical students were then asked to think aloud as they worked through the items. McGuire reported that comparing the experts' classifications to the students' verbal reports helped her to align better the objectives of instruction with their assessment. However, it is unclear exactly how the verbal reports were ex-

amined and how the information in them was used to develop the test. As in the previously described studies, no specific guidance is given on how verbal reports of thinking can be used in test development.

Connolly and Wantman (1964) reported a study intended to improve that of Bloom and Broder (1950). Nine subjects were asked to report all the thoughts that might cross their minds as they worked on a set of 25 verbal analogy items. The adequacy of the subjects' thinking was compared with a model of ideal thinking on the items. The model was general rather than item specific and listed 67 behaviors representing quality of thinking (e.g., "Justifies choice with incorrect logic," and "Reads stem and immediately states keyed response"). The attempt to use a normative model of reasoning to judge verbal reports is significant for testing critical thinking, because critical thinking theory is a normative theory, that is, not merely a theory of thinking, but a theory of how to think *well*, namely, in accord with a set of standards that define how to think well (Ennis, 1981). However, Connolly and Wantman (1964) provided insufficient detail of what their ideal model is, where it originated, and how it was used. They concluded that their technique was useful for pretesting items, but did not mention a role for the technique in test validation. They suggested that the method might be used to determine whether some students are penalized for having too much knowledge. This suggestion is particularly relevant to multiple-choice critical thinking testing because of the allegations that it is knowledge recall and not critical thinking that these tests examine.

A study by Schuman (1966) provides the most detailed procedure for using the information in verbal reports of examinees' thinking. One thousand factory workers and cultivators in East Pakistan completed an attitude survey and then reported on the reasons for their responses to 10 randomly selected items. Their reasons were rated on a scale from 1, given to a reason that was clear and led to an accurate prediction of the chosen response, to 5, given to a reason that was unclear and could not be used to predict the answer chosen. Average scores were calculated for each item across all individuals and were used to indicate the subjects' understanding of that item: the lower the score the better the understanding. Schuman's procedure for quantifying and using the information in examinees' verbal reports marks an important difference in approach from previous studies, but the approach is not without problems. For instance, it was possible for examinees' reasoning to express clearly a *mis*understanding of an item, yet reliably lead to predictions of their responses. Their reasoning would be assigned 1, and the item would receive endorsement from a low average score, even though subjects did not understand it.

In sum, the research on using verbal reports of thinking for developing and validating tests is not adequate. Few systematic attempts have been made to use the information contained in verbal reports to quantify the quality of examinees' thinking and to accumulate the ratings of quality for each item across individuals. Furthermore, little emphasis has been placed on developing normative models against which reports of thinking might be compared. Without such models, the choice of keyed answers cannot be justified except by some atheoretical consensus procedure. Verbal reporting techniques have

been used rather casually in designing multiple-choice tests, with the possible exception of Schuman's study.

A Method for Eliciting and Using Verbal Reports of Thinking

To use systematically verbal reports of thinking in the development and validation of multiple-choice critical thinking tests, I am proposing a nine-step procedure as follows:

1. developing a set of critical thinking principles that items would serve to test;
2. devising initial trial items to test the principles;
3. developing normative models of thinking on each item;
4. developing an interviewing methodology for eliciting trustworthy reports of thinking;
5. collecting verbal reports of thinking on items;
6. scoring verbal reports using the normative models of thinking as a standard, and scoring answers chosen according to the answer key;
7. comparing thinking on items to performance on them and systematically accumulating these comparisons across examinees for each item;
8. modifying suspected items;
9. retrying the test in accord with steps 5 to 8.

Following is a demonstration of each of these steps in the development of a multiple-choice credibility judgment test.

Developing Principles of Observation Appraisal

In order to assess critical thinking and to compare the quality of examinees' verbal reports of thinking with their choice of answers, we need some justified standard for rating quality of thinking, that is, a set of norms against which thinking can be judged. The particular test being used as an illustration here, the Test on Appraising Observations (Norris & King, 1983), was designed to assess ability to appraise reports of observations. Therefore, to judge the quality of examinees' thinking on this test, a comprehensive and defensible set of principles for appraising observations is needed. Several sources provide such principles. For example, principles based on observers' conflict of interest, on being asked leading questions, and on the skill of the observer find support from longstanding judicial practice (Brooks, 1983). The principle that being asked leading questions lowers the credibility of answers is also supported by psychological research on eyewitness testimony (Lindsay, Wells, & Rumpel, 1981; Loftus, 1979; Wells, Ferguson, & Lindsay, 1981; Yarmey, 1979). Several principles are supported by commonsense psychology. For example, it makes sense to think that an observation is more believable when made by a more alert observer.

Ennis (1981) has compiled a comprehensive set of principles from a variety of such sources and Norris (1984) and Norris and King (1984) have extended Ennis's set. A sample of such principles is contained in Table 1. The principles do not provide either necessary or sufficient conditions for observation statements to be credible. They suggest conditions to consider. To the extent that these conditions are satisfied, an observation is worthy of belief, and to the extent they are not satisfied, it is worthy of doubt. In any given case, principles

Table 1
Illustrative Principles for Appraising Observations Selected from a Comprehensive Set by Norris and King (1984)

I. Observation statements tend to be more credible than inferences based upon them.
II. An observation statement tends to be credible to the extent that the *observer:*
    1. is alert to the situation and gives his or her statement careful consideration;
    2. has no conflict of interest;
    3. is skilled at observing the sort of thing observed.
III. An observation statement tends to be credible to the extent that the *observation conditions:*
    1. provide a satisfactory medium of observation;
    2. provide sufficient time for observation;
    3. include adequate instrumentation, if an instrument is used.
IV. An observation statement tends to be credible to the extent that the *observation statement:*
    1. is corroborated;
    2. is made close to the time of observing;
    3. is not given in response to a leading question.

may compete, and there are no strict procedures for arbitrating such difficult cases. Good judgment and experience are needed.

Devising Initial Trial Items

Once a set of principles is compiled and defended, items can be constructed to test knowledge of them. The construction requires that situations calling for judgments using the principles be devised, and that the conclusions implied by the principles in these situations be determined. These conclusions then serve as a key, and examinees' knowledge of the principles can be inferred from whether their conclusions match those in the key.

Based on this general scheme, initial trial items were written. They were in multiple-choice format and each was written to test ability to use one principle of observation appraisal. Fifty items were cast in the context of two stories: Part A, the story of a traffic accident, and Part B, the story of a river exploration. The story context was used in order to provide examinees sufficient context for deciding which principle of observation appraisal should be applied.

Each item presented two conflicting reports of observations made by characters in the stories and required examinees to decide which, if either, of the reports was more credible. Comparative judgments were requested because, without the wealth of information available in real situations, usually it is not possible to make absolute judgments about the credibility of single observation statements (Ennis, 1988; Norris & Ennis, 1989). Thus for each item the three possible answer choices were: the first statement is more credible, the second statement is more credible, and the statements are equally credible. Multiple-choice items often have more than three alternatives to reduce the likelihood of guessing correctly, but the above three are the only logically obvious ones given the task.

Developing Normative Models of Thinking for Items

The principles of observation appraisal were used to develop a normative model of thinking on each item, that is, a model that instantiates one of the principles of observation appraisal and uses the principle to support a judgment of credibility. As an example, consider Item 3 from Part A, the traffic accident story:

> A policewoman has been asking Mr. Wang and Ms. Vernon questions. She asks Mr. Wang, who was one of the people involved in the accident, whether he had used his signal.
> Mr. Wang answers, "Yes, *I did use my signal.*"
> Ms. Vernon had been driving a car which was not involved in the accident. She tells the officer, "*Mr. Wang did not use his signal.* But this didn't cause the accident."

Examinees were to choose which, if either, of the emphasized statements was more credible. To reason through this question correctly, an examinee first needs to be able to derive from the text the relevant information about Wang's and Vernon's involvement. The text is simple enough that most high school students should have no difficulty with this aspect of the task. Second, an examinee must retrieve from background knowledge the relevant facts that not using a turn signal can cause an accident and that being held responsible for an accident can cause one trouble with the law. Again, high school students would have such common knowledge. Finally, an examinee has to infer that the possibility that one's own testimony can place one in trouble with the law is an accuracy-reducing factor, and thus that Wang is less credible than Vernon.

The normative model of thinking developed for this item is as follows:

> Mr. Wang was involved in the accident, but Ms. Vernon was not involved. Mr. Wang is less credible because his involvement would give him reason to say he used his signal even if he did not. Wang is in a conflict of interest. People in a conflict of interest, that is, people who have something to gain by events being cast as they described them, tend to be less credible than those who are not in such a situation.

This normative model instantiates the principle of conflict of interest by showing how Wang is in such a conflict, and then uses the principle to support the judgment that Wang is less credible than Vernon.

The reasoning required for Item 3 generalizes to the following three steps that are required for sound reasoning on all the test items:

1. cite the relevant facts in the text that can be used to compare the credibility of the emphasized statements;
2. use these facts together with relevant background knowledge to make a comparative evaluation of the credibility of the statements;
3. show how the evaluation is based on an appropriate general principle that subsumes the relevant facts in the text under a general accuracy-reducing or accuracy-increasing factor.

Each item on the test was analyzed according to this scheme, normative models of reasoning for each item developed, and the models used in a

manner described subsequently to rate the quality of examinees' verbal reports of thinking.

Developing an Interviewing Methodology

An interviewing methodology was designed to elicit examinees' reasoning while they worked through items on the observation test. In accord with widely advocated practices for collecting verbal reports as data (Afflerbach & Johnston, 1984; Ericsson & Simon, 1980, 1984; Larkin & Rainard, 1984; Loftus, 1979; Norris, 1991) the interviews were as nonleading as possible. The initial directive to examinees was: "As you do each question tell me all you can about what you are thinking while you are picking your answer." At this stage, interrupting an examinee's verbal reporting was permitted only to clarify the ambiguous referent of a pronoun or to point to obvious reading errors.

It is important to obtain records of reasoning that are as complete as possible. To fulfill this aim it is necessary sometimes to probe beyond the initial instruction to think aloud, while being careful not to rush examinees by prematurely cutting off their reasoning (Larkin & Rainard, 1984) or to endorse or criticize particular reasoning attempts. Even in such follow-up stages, it is desirable for probing to be as nonleading as possible, by merely echoing examinees' reported thoughts or by asking them to explain a little more what they have said.

Some researchers who use this method think that some training is needed for subjects to give useful verbal reports of their thinking. To this end, such techniques as using demonstration tapes (Bridge & Winograd, 1982), practice verbal reporting sessions (Kavale & Schreiner, 1979), practice examples (Larkin & Rainard, 1984), and prestudy introspection (Johnston & Afflerbach, 1983) have been tried. However, it is not known what effect, if any, such practice has on the verbal reports that subjects subsequently give (Afflerbach & Johnson, 1984). For this reason, and because of the amount of time that we asked subjects to devote to the actual data gathering phase, training sessions were not used in the research reported here.

To meet the above goals and constraints, the three-stage interview methodology given in Appendix A was adopted. In Stage One the interviewer informs examinees of the general purpose of the interviews and of their role in them, especially of the fact that they will be asked some questions about their thinking to be answered verbally while taking a test.

In Stage Two the interviewer asks examinees to tell all they can about what they are thinking while choosing their answers. Interruptions to examinees' reports are permitted only to probe for ambiguous references and for obvious reading errors (see II.2.1 and II.2.2). If examinees ask for additional information or for reasons, the interviewer must respond as in II.3.1, "You can only go by what is written," or II.3.2, "You can decide only according to what is said and what you know." If examinees seek feedback on their progress it must not be given (II.4.3).

Stage Three begins after examinees have chosen their response and have finished reporting on their thinking in response to the directive in II.1. The interviewing procedure for this stage consists of a set of condition-action pairs that guide the actions of the interviewer in accord with the satisfaction of conditions defined by a combination of examinees' answer choices and the

content of their verbal reports. Each action is designed to match a particular combination by being as nonleading as possible and still having a chance to elicit useful information. For any given item and examinee response, only one numbered step applies. The strategies listed under III.4 guide the interviewer when it is not clear how to proceed given an examinee's report.

Stage Three proceeds as follows. Starting with III.1 and proceeding sequentially, the interviewer checks whether the condition (described by the phrase following the *if*) is satisfied. Sometimes the condition is complex, that is, consists of subconditions, so all subconditions must be met for the condition to be satisfied. If a condition is not satisfied, the interviewer proceeds to the next numbered step. If a condition is satisfied, the interviewer carries out the indicated action (described by the phrase following the *then*). When the signal *then proceed to the next item* is reached the interview reverts to Stage Two.

Collecting Verbal Reports of Thinking

Our experience has shown that a multiple-choice test takes about four times as long in a verbal reporting format as in a paper-and-pencil format. Therefore, it is too tiresome for students to be interviewed on a whole test designed for a regular class period when taken in paper-and-pencil format. However, if there are story lines in a test and items are linked thematically, as is the case with the Test on Appraising Observations, or if the items are in a logical sequence for another reason, a student could not be interviewed on a later section without having completed previous sections. A procedure for interlacing interviewing and paper-and-pencil formats is needed.

Subjects whose verbal reports of thinking were to be collected were divided randomly into halves; one half was assigned to take Part A of the test and the other half to take Part B. Each part of the test was divided into two sections representing logical breaks in the story line: Part A, Section I (Items 1-15) and Section II (Items 16-28); Part B, Section I (Items 29-37) and Section II (Items 38-50). Half of the subjects assigned to a part were randomly chosen to be interviewed on Section I of that part and to complete Section II in paper-and-pencil format; the other half was to write Section I and be interviewed on Section II.

An interviewer dealt with two students at a time working on the same part, either A or B, of the test. First, both students were told how the testing would proceed. The purpose of the study had already been explained in meetings with students in their classes. The student who was to write Section I of the part was taken to a place to work alone, instructed to stop when finished the last item of Section I (either Item 15 or 37), not to read on, and to wait to be called by the interviewer. The other student was interviewed on Section I according to the methodology described in Appendix A. When the interview was over, the students exchanged places. The student who wrote Section I was interviewed on Section II and the student who was interviewed on Section I wrote Section II. Students marked their answers to all questions on a standardized answer sheet, and students' verbal reports were tape-recorded.

The results of two such interviewing stages, for two sequential versions of the Test on Appraising Observations (Versions B and C), are described and compared in this article. For version B, 52 grades 10 and 11 students from seven small rural schools took part. For Version C, 45 students from four

Table 2
Numbers of Examinees Giving Verbal Reports on Each Test Section for Versions B and C of the Test on Appraising Observations

Version	Part A		Part B	
	Sect. I	Sect. II	Sect. I	Sect. II
B	13	14	13	12
C	13	12	10	10

different rural schools were involved. Students were selected randomly from class lists and given the opportunity to participate. None refused. Verbal reports for each test section were collected on the number of students shown in Table 2.

Scoring Verbal Reports of Thinking and Answer Choices

Performance scores were assigned to answer choices for each item according to whether they agreed (1) or disagreed (0) with a key based on the set of principles illustrated in Table 1.

Verbal reports were transcribed verbatim and analyzed. Each student was assigned a *thinking score* for each item according to the rating scale in Table 3. The rating scale is based on the normative models of thinking and assigns scores according to how closely students' verbal reports approximate the models. Students were not privy to the rating scale. Thus, the amount of detail and generality provided in students' verbal reports was entirely up to them. The aim was to determine what students would say when left to their own resources, since this best simulated the paper-and-pencil text administration. Table 4 shows how the rating scale would work for Item 3.

For illustration, here are the verbatim transcriptions of the verbal reports of two students working on Item 3:

Student A ... ah ... ah, Mr. Wang, like he probably didn't, like you know, it was just, he probably thought he used his signal, but really didn't. And Miss Vernon, she was watching, so she'd be able to

Table 3
Rating Scale for Assigning Thinking Scores

Rating	Basis of Evaluation
1	The examinee indicates a criterion by which a correct comparison of the emphasized statements might be made, or the examinee uses a criterion in comparing the two statements but does not explicitly indicate what the criterion is.
2	The examinee indicates a criterion and correctly uses it to compare the emphasized statements.
3	The examinee indicates a criterion, correctly uses it to compare the emphasized statements, and justifies the comparison on the basis of a general principle.
0	The examinee does none of the above, does not think critically in some other way, or does not respond.

Table 4
Assigning Thinking Scores for Item 3

Rating	*Basis of Evaluation*
1	The examinee points out that Mr. Wang was involved in the accident.
2	The examinee points out that Mr. Wang was involved in the accident and compares Mr. Wang's involvement to Ms. Vernon's being a bystander.
3	The examinee points out that Mr. Wang was involved in the accident, compares this with Ms. Vernon's noninvolvement, and shows that these facts put Wang in a position of people who stand to profit or lose depending upon what they say.
0	The examinee does none of the above, does not think critically in some other way, or does not respond.

tell from back if he was using it or not. Right? Being the case, so, I'd tend to believe Vernon.

Student B I would say that he did use his signal because anybody who's in the car ... coming up to an intersection or anything ... he, he usually knows what he's doing. So I'd be more inclined to believe the first.

Student A chose the keyed response and Student B an unkeyed one. Neither student thought critically on the item, however, and both were assigned zero for thinking scores. The criteria used by both students for comparing Wang and Vernon were unsound. Student A's reasoning that Wang just "thought he used his signal," but Vernon would "be able to tell from back" because "she is watching" is arbitrary. There is no information in the item to justify the claims about either Wang or Vernon. It is used to rationalize a choice of answer rather than to justify it. Given the information in the item, it is just as reasonable to say that Wang would "be able to tell from inside that he used his signal" and that Vernon just "thought he did not use his signal." Student B's reasoning fails to distinguish between Wang's *knowing* what he had done and his *reporting accurately* what he had done, and to allow that someone in another car can know what another driver is doing. The students' performance scores and thinking scores for Item 3 are as in Table 5.

For each student, there was a performance score for each item taken and, for each item done in the interview, a thinking score also. Performance and thinking scores were conceptually and empirically independent: there is little difficulty in seeing that although Student A chose the keyed answer, the student's justification was poor; across the two students, performance and

Table 5
Two Students' Performance and Thinking Scores for Item 3

Student	*Performance Score*	*Thinking Score*
A	1	0
B	0	0

thinking scores are uncorrelated. Thinking scores were assigned without reference to answer chosen. When appropriate, students were given credit for using correctly criteria other than those assumed when designing the items. That is, the ideal model of thinking for each item was used as a guide, but it was possible for reasoning to be rated highly according to general standards of critical thinking even though it was outside the bounds of the specific normative model on which this test is based. For instance, an examinee might reject a factual presupposition built into an item and reason perfectly acceptably from an alternative presupposition to other than the keyed response. In such a situation, the thinking score would be high, but the performance score would be zero.

Comparing Performance and Thinking Scores and Judging Items

A central feature of the test development and validation methodology is the role that the relationship between performance and thinking scores plays in the evolution of a test. The aim is to adjust items, or to write new ones, so that the relationship between quality of thinking and quality of performance is strong for each item. The relationship was calculated using point biserial correlation coefficients between performance scores and thinking scores for each item across subjects sampled. Consider the following two sets of ordered pairs of thinking and performance scores for Items 18 and 38 in Version B of the test:

{18 | (2,1)(0,0)(0,0)(2,1)(2,1)(2,1)(2,1)(0,0)(2,1)(2,1)(1,1)(2,1)(2,1)(2,1)};

{38 | (0,1)(0,0)(0,1)(0,0)(0,0)(0,0)(1,0)(0,0)(0,1)(0,0)(1,0)(1,1)}.

For Item 18 all the thinking/performance combinations support the validity of the item. That is, for each subject, choosing the keyed answer is associated with thinking well and choosing an unkeyed answer is associated with thinking poorly. This relationship is reflected in a high thinking/performance point biserial correlation of 0.96. Comparing Item 38, for which there are five mismatches between thinking and performance scores (either combinations [1,0] or [0,1]), the computed point biserial correlation is 0.33. This evidence suggests that Item 18 is suitable as it stands, but that Item 38 needs modification. An examination of traditional item statistics, difficulty levels, and item/test biserial correlations also suggest Item 38 needs modification. For Items 18 and 38, respectively, we have: difficulties of 0.56 and 0.46; item/test biserials of 0.45 and 0.21.

However, this correspondence between traditional indicators of item validity and the thinking/performance indicator recommended here did not obtain for all items. For instance, Item 5 on Version B had a very high thinking/performance point biserial of 0.89, but an item/test biserial of 0.19. Placing more emphasis on the thinking/performance relationship for items is an attempt to meet the challenge from theorists of critical testing that traditional measures of internal consistency (KR-20 reliability and item/test biserials) may not suit critical thinking testing, because critical thinking is a heterogeneous concept. The approach here was to treat the item/test biserial as a dichotomous indicator—only if the item/test biserial for an item was negative was this indicator used to call an item into question.

Modifying Suspected Items

As described in the previous section, item level statistics, such as item difficulty levels, item/test biserial correlations, and thinking/performance correlations were used to identify problematic items in experimental versions of the test. Most weight was given to the thinking/performance correlation. There were 19 items on Version B with thinking/performance correlations <0.5 (rounded to one decimal place). This figure was taken as the lower bound of acceptability, and reasons were sought for these low indexes by examining the transcribed interviews and asking the following questions:

1. Do examinees understand the task in the intended way; or are the instructions unclear, is the item systematically misleading, or is the assumed background knowledge outside their knowledge store?
2. Are examinees thinking critically to arrive at other than the keyed response and, if so, is it the keyed response or is it the item that should be changed?
3. Are examinees thinking uncritically yet arriving at the keyed answer and, if so, are there clues in the text that can be eliminated?

For Version B, all but four of the 19 suspect items, items 12, 17, 22, and 23, were altered on the basis of information contained in the verbal reports. Alterations included single word changes, the addition and deletion of information, and changes to the directions. Items 12, 17, 22, and 23 were not altered because there was no clear basis for alteration contained in the verbal reports and because it was expected that changes to neighboring and thematically related items would improve them also.

An example of changes made to the directions serves to illustrate how the thinking/performance correlations and the verbal reports were used together. Recall that Part A of the test was set in the context of a traffic accident. The Version B directions listed the names of all characters and what they were doing at the time of the accident. With several characters in the story in different roles, this list was provided to help examinees keep them straight. But the verbal reports showed that many examinees looked for direct answers to questions in this information. In one item, for example, characters gave conflicting reports about how many cars were at the intersection when the accident occurred. One character was more alert to the traffic and hence a more credible source of information about the number of cars at the intersection. Coincidentally, that character reported seeing the same number of cars *at the intersection* (3) as the directions mentioned were *involved in the accident*. The verbal reports showed that several examinees did not consider the characters' alertness, but merely counted the number of cars mentioned in the directions and chose the keyed answer as a result. These students thought uncritically, because they equated the number of cars at the intersection with the number involved in the accident, but chose the keyed response. It is the possibility of just this sort of occurrence with multiple-choice tests that advocates of critical thinking find most disturbing.

Retrying the Modified Test

When the 19 suspected items were examined and 15 of them revised, the new version (Version C) was tried by collecting verbal reports of thinking on it, calculating thinking/performance correlations, and so on. Figure 1 displays a comparison of the thinking/performance correlations for Versions B and C.

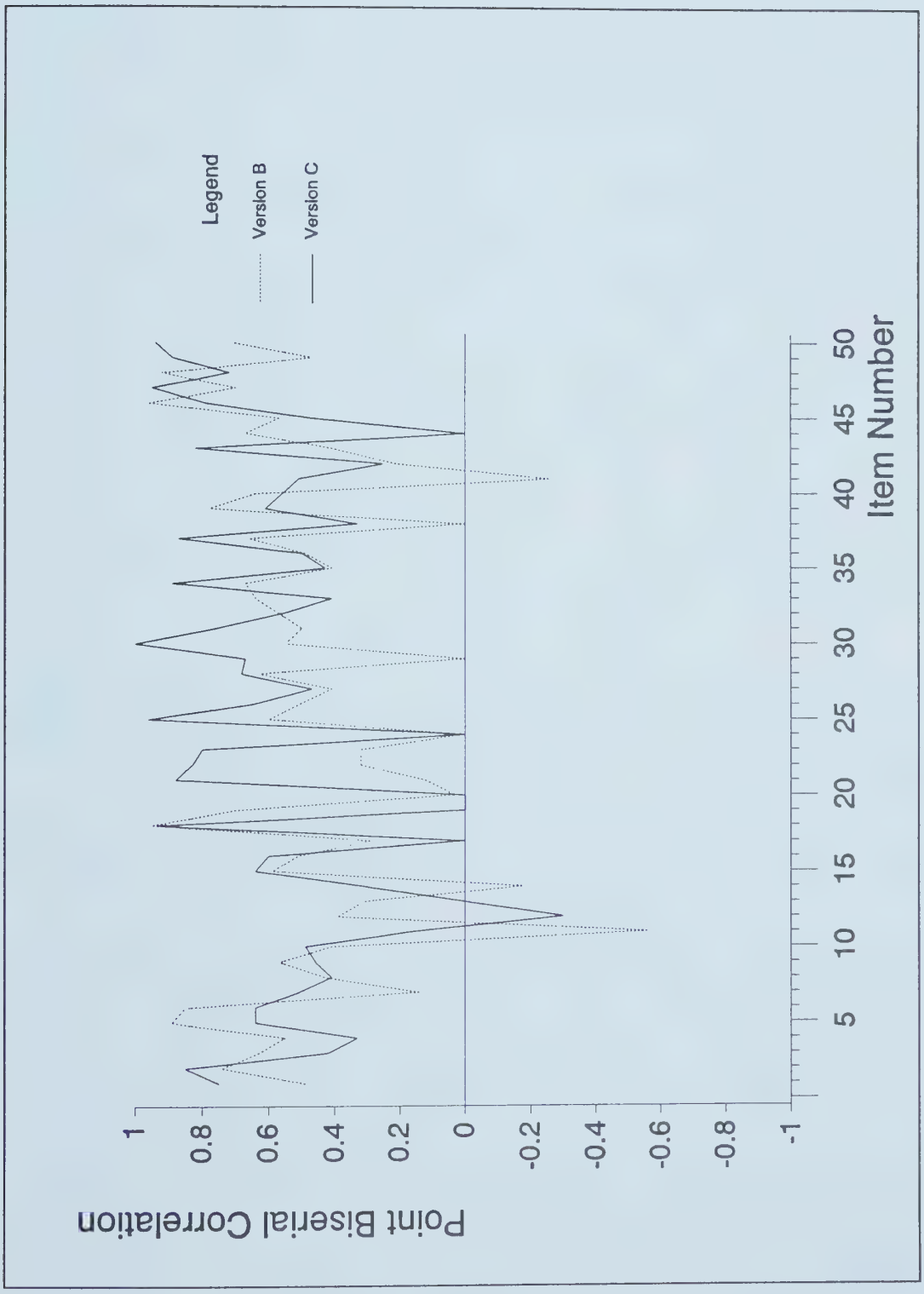

Figure 1. Thinking/performance correlations by item for Versions B and C.

There was a general increase in thinking/performance correlations for Version C: 29 items in Version C had larger thinking/performance correlations than Version B; 20 items had smaller correlations; and one correlation was equal for the two versions. The most consistent increases were made in the range of Items 21 to 37. Most telling, there was still one negative correlation, for Item 12, and six zero correlations. For three of the zero correlations, all examinees answered the item correctly and had thinking scores greater than zero. The items were too easy for this sample. In the other three, all examinees received zero for thinking and up to four students chose the keyed answers. These items were too difficult, but, many would argue, should remain on the test because they represent critical thinking principles that subjects should know.

Other items on Version C clearly needed further revision and trial. The interviews indicated for items 3, 4, and 8 a common error in students' thinking that was largely responsible for the low thinking/performance correlations. In each item, several students made their choice solely on the basis of who was the driver of a car on the grounds that a driver would be more alert to the traffic than a pedestrian or passenger. For example, in Item 3 on Version C, Mr. Wang was a driver and Ms. Vernon "watched the accident happen." Several students chose Mr. Wang's statement as more credibile because he was a driver and thus more alert to road conditions. These students were assigned thinking scores of 1 for recognizing this important factor, even though they neglected the more important fact that Wang was in a conflict of interest. Thus they chose an unkeyed response but thought fairly well, contributing to a lowered thinking/performance correlation. The obvious change was to make both Wang and Vernon drivers, as is the case in the version of Item 3 cited previously as an example. Items so modified require retrial following the procedures followed in modifying Version B.

Discussion and Conclusions

Verbal reports of thinking on trial items of multiple-choice critical thinking tests seem useful and important from at least three perspectives. First, existing multiple-choice critical thinking tests are often criticized because they give no direct evidence of the process of thinking that leads to examinees' choices of answers. The criticism is justified and important, because evidence on the thinking processes used to answer items has a direct bearing on construct validity. If examinees choose keyed answers but they do not think critically, or choose unkeyed answers but think critically, then the test is not valid.

Second, multiple-choice critical thinking tests could serve several useful purposes if they were valid. They are useful when testing groups of students, when time for grading is at a premium, and when knowledge of several criteria and principles of critical thinking is being tested. That is, multiple-choice critical thinking tests can fill a particular set of roles, though they probably cannot serve all the needs of critical thinking assessment. For instance, multiple-choice tests probably cannot assess ability to think critically in complex situations where solutions are radically underdetermined by available information; they cannot test for the ability to orchestrate many critical thinking abilities; and they cannot test for critical thinking dispositions.

Third, direct evidence on the information, strategies, and principles that examinees use to answer items on existing multiple-choice critical thinking

items would be directly relevant to the construct validity of those tests. In the context of test design, direct evidence from examinees' verbal reports can be used to discard, modify, and retain items as done in this study. In the context of test use, the test would inherit the weight of this direct evidence gathered in the design context, even though in the use context no direct evidence would be gained on the causes of performance. In addition to being useful in test development, the verbal reporting procedure would be useful for evaluating the claims that currently available multiple-choice critical thinking tests do not measure critical thinking (Ennis & Norris, 1990; McPeck, 1981; Petrie, 1986; Whimbey, 1985). For instance, the procedure could be adapted for studying the validity of the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980) and the Cornell Critical Thinking Test Levels X and Z (Ennis & Millman, 1985), two of the oldest and most widely used multiple-choice critical thinking tests.

A number of possible outcomes emerge from such studies. First is the possibility that the criticisms of these tests would be sustained. This would be unfortunate, because it would cast doubt on much of the research on the effectiveness of critical thinking instruction that used the tests as criterion measures. Another possible result is a clearer understanding of the range of applicability of existing critical thinking tests. Many multiple-choice critical thinking tests are advertised for use over wide age ranges. For example, the Cornell Critical Thinking Test Level X is said to be usable from grade 4 to grade 14 and the Watson-Glaser Critical Thinking Appraisal from grade 9 up. These ranges include individuals who differ widely in intellectual sophistication, empirical knowledge of the world, and test-taking experience, so it is quite plausible that the tests are not equally suitable across the entire ranges. A study of the verbal reports of thinking of examinees from different groups would provide direct evidence on the range of suitability of each test. A possible consequence is that the range for each test will need to be narrowed. Another possibility is that answer keys might be adjustable to match more suitably the critical thinking of examinees of different age levels.

The verbal reporting approach itself needs further study using different types of test content, different elicitation procedures, and different types of examinees. It is known, for instance, that different questioning procedures can affect differently the reports eyewitnesses give of events (Loftus, 1979). Can different approaches to eliciting verbal reports of thinking alter what examinees report? One study (Norris, 1990) shows that verbal reports of high school students' thinking are not easily altered by asking them more leading questions, and that performance scores while giving verbal reports are the same as performance scores in a paper-and-pencil test taking situation. This is promising. But would verbal reporting work as well with other groups and in other testing situations?

Indirect evidence for the value of the approach can be found by examining standard indicators of test validity and reliability. The Cornell Critical Thinking Test (Ennis & Millman, 1985) and the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980) provide criteria for judging the validity of the Test on Appraising Observations and hence of indirectly judging the effectiveness of the verbal reporting technique. Those tests have 71 and 80

items respectively, whereas the Observations test has 50 items. This would tend to make the Observations test less reliable. Comparison of the mean KR-20 reliabilities for the three tests computed from the performance of four groups of senior high school students from Southern Ontario (N=100, 101, 106, 287) shows that the reliability of the Test on Appraising Observations is only slightly less than the Cornell (0.69 vs. 0.72). The Watson-Glaser reliability of 0.80 is about what would be predicted by increasing the number of items from 50 to 80 (see Lord & Novick, 1968, Fig. 5.10.1, p. 113). Thus reliability does not seem to have been sacrificed to validity in using verbal reports and focusing on item validity to develop the observations test.

It was also expected that the correlation of the Observations test with the Cornell test should be higher than with the Watson-Glaser test. The rationale for this expectation is that the Cornell test includes a section of 24 items on judging the credibility of reports of observations, but the Watson-Glaser has no items on judging credibility. Using the data on the same four groups of Southern Ontario students, the average correlation between the Observations and Cornell test was 0.48 ($p<.01$), while the average correlation between the Observations and the Watson-Glaser test was 0.31 ($p<.01$), in line with predictions. Also in line with predictions, for a group of 172 grades 10, 11, and 12 students with a wide range of academic ability attending a large metropolitan school in the Atlantic provinces, the correlation between the Observations and Cornell tests was 0.62 (Bonferroni adjusted $p<.05$), and the correlation between the Observations and Watson-Glaser tests was 0.35 (Bonferroni adjusted $p<.05$).

The verbal reporting methodology thus led to a test that meets several expectations. This is evidence for the value of the methodology. Therefore, eliciting verbal reports of examinees' thinking on multiple-choice critical thinking tests, comparing the information in the verbal reports to examinees' choices of answers, and using these comparisons systematically to weigh and balance the evidence for the quality of items seems a valuable approach for developing and validating such tests. In particular, it is a way to satisfy two competing desires of many proponents of critical thinking: the desire to use multiple-choice critical thinking tests because they provide a convenient way to gather information on students' critical thinking and the desire to have information on the process, not just the products, of that thought.

Acknowledgments

This research was supported by grants from the Social Sciences and Humanities Research Council of Canada (Grant Nos. 418-81-0781 and 410-89-0052).

I thank Linda M. Phillips and an anonymous reviewer for helpful comments.

References

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior, 41*(4), 307-322.

Anastasi, A. (1988). *Psychological testing*. New York: Macmillan.

Bloom, B.S., & Broder, J.L. (1950). *Problem-solving processes of college students*. Chicago: University of Chicago Press.

Bridge, C., & Winograd, P. (1982). Readers' awareness of cohesive relationships during cloze comprehension. *Journal of Reading Behavior, 14*, 299-312.

Brooks, N. (1983). *Police guidelines: Pretrial eyewitness identification*. Ottawa: Law Reform Commission of Canada.

Connolly, J.A., & Wantman, M.J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement, 1*, 59-64.

Cronbach, L.J. (1971). Test validation. In R.L.Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S., Schneider, L.M., & Roth, D.L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tasks. *Journal of Educational Measurement, 23*, 13-32.

Ennis, R.H. (1981). Rational thinking and educational practice. In J. Soltis (Ed.), *Philosophy and education . Eightieth yearbook of the National Society for the Study of Education, Part 1* (pp. 143-183). Chicago: NSSE.

Ennis, R.H. (1987). A taxonomy of critical thinking dispositions and abilities. In J.B. Baron & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York: Freeman.

Ennis, R.H. (1988). Testing teachers' competence, including their critical thinking ability. In B. Arnstine & D.G. Arnstine (Eds.), *Philosophy of Education 1987* (pp. 413-420). Normal, IL: Philosophy of Education Society.

Ennis, R.H., & Millman, J. (1985). *Cornell critical thinking test, level X*. Pacific Grove, CA: Midwest Publications.

Ennis, R.H., & Norris, S.P. (1990). Critical thinking assessment: Status, issues, needs. In S. Legg & J. Algina (Eds.), *Cognitive assessment of language and math outcomes* (pp. 1-42). New York: Ablex.

Ericsson, K.A., & Simon, H.A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.

Ericsson, K.A., & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R.O. Freedle & R.P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Johnston, P., & Afflerbach, P. (1983, December). *The construction of main idea in reading comprehension*. Paper presented at the annual meeting of the National Reading Conference, Austin, Texas.

Kavale, K., & Schreiner, R. (1979). The reading processes of above average and average readers: A comparison of the use of reading strategies in responding to standardized comprehension measures. *Reading Research Quarterly, 15*, 102-128.

Kropp, R.P. (1956). The relationship between process and correct item responses. *Journal of Educational Research, 49*, 385-388.

Larkin, J.H., & Rainard, B. (1984). A research methodology for studying how people think. *Journal of Research in Science Teaching, 21*(3), 235-254.

Lindsay, R.C.L., Wells, G.L., & Rumpel, C.M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*, 79-89.

Loftus, E.F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McGuire, C. (1963). Research in the process approach to the construction and analysis of medical examinations. *National Council on Measurement in Education Yearbook, 20*, 7-16.

McPeck, J. (1981). *Critical thinking and education*. New York: St. Martin's Press.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

National Assessment of Educational Progress. (1985). *The reading report card: Progress toward excellence in our schools: Trends in reading over four national assessments, 1971-1984*. Princeton, NJ: Educational Testing Service.

Norris, S.P. (1984). Trying eyewitness testimony, validating tests, and teaching critical thinking. In R.E. Roemer (Ed.), *Philosophy of Education 1983* (pp.179-189). Normal, IL: Philosophy of Education Society.

Norris, S.P. (1989). Can we test validly for critical thinking? *Educational Researcher, 18*(9), 21-26.

Norris, S.P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement, 27*, 41-58.

Norris, S.P. (1991). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In J.F. Voss, D.N. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 451-472). Hillsdale, NJ: Erlbaum.

Norris, S.P., & Ennis, R.H. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Midwest.

Norris, S.P., & King, R. (1983). *Test on appraising observations*. St. John's, NF: Institute for Educational Research and Development, Memorial University of Newfoundland.
Norris, S.P., & King, R. (1984). Observation ability: Determining and extending its presence. *Informal Logic, 6*(3), 3-9.
Paul, R. (1985). Bloom's taxonomy and critical thinking instruction. *CT News, 3*(5), 3-6.
Petrie, H. (1986). Testing for critical thinking. In D. Nyberg (Ed.), *Philosophy of Education 1985* (pp. 3-19). Normal, IL: Philosophy of Education Society.
Resnick, L.B. (1987). *Education and learning to think*. Washington, DC: National Academy Press.
Schuman, H. (1966). The random probe: A technique for evaluating the validity of closed questions. *American Sociological Review, 31*, 218-222.
Spiro, R.J., Vispoel, W.L., Schmitz, J.G., Samarapungavan, A., & Boerger, A.E. (1987). Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In B.C. Britton & S. Glynn (Eds.), *Executive control processes* (pp. 177-199). Hillsdale, NJ: Erlbaum.
Watson, G., & Glaser, E.M. (1980). *Watson-Glaser critical thinking appraisal*. Cleveland, OH: Psychological Corporation.
Wells, G.L., Ferguson, T.J., & Lindsay, R.C.L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*, 688-696.
Whimbey, A. (1985). You don't need a special "reasoning" test to implement and evaluate reasoning training. *Educational Leadership, 43*(2), 37-39.
Yarmey, A.D. (1979). *The psychology of eyewitness testimony*. New York: Free Press.

Appendix A
Interviewing Methodology for Eliciting Examinees' Reasoning on Multiple-Choice Tests of Critical Thinking (with Examples Based on the Test on Appraising Observations)

Stage One

I.1 Inform examinee of purpose of interview: to find out what he or she is thinking while choosing answers to questions on the test.

I.2 Inform examinee of his or her role: to respond as completely as possible to the questions asked on the test.

Stage Two

II.1 Interviewer says to examinee:
"As you do each question tell me all you can about what you are thinking while you are picking your answer."

II.2 Interviewer can interrupt examinee's narrative only to:

II.2.1 probe for ambiguous reference of demonstratives or third person pronouns by saying: "Could you tell me what you mean by ...?"
Example: When doing Item 3 the examinee says: "This driver here is more believable." Probe immediately: "Could you tell me what you mean by 'this driver here'?"

II.2.2 probe for obvious reading mistakes by saying:
"Did you read ...?" (Do not endorse answers.)
Example: When doing Item 3 the examinee reads: "Mr. Wang did use his signal" for "Mr. Wang did not use his signal." Probe immediately: "Did you read 'Mr. Wang *did* use his signal'?"

II.3 Interviewer can respond to examinee's questions only as follows:

II.3.1 If examinee probes for facts, say: "You can only go by what is written."
Example: When doing Item 3 the examinee asks: "Was Vernon close enough to see Wang's car?" Answer only: "You can only go by what is written."

II.3.2 If the examinee probes for reasons, say: "You can decide only according to what is said and what you know."

Example: When doing Item 3 the examinee asks: "How did Vernon know what caused the accident?" Answer only: "You can decide only according to what is said and what you know."

II.4 General cautions:

II.4.1 Do not begin to speak immediately after the examinee stops; give the examinee a few seconds to continue.

II.4.2 Do not cut off examinee's reasoning by, in any way, signalling that enough has been said, even though the examinee might seek such signals.

II.4.3 Do not endorse or criticize the examinee's fact-finding or reason-giving.

Stage Three

III.1 If "neither" is the answer chosen, then say: "So you believe neither is more believable?" Wait for examinee to respond and *then proceed to next item.*

III.2 If the item is an inference vs. observation question, that is, a question testing Principle I in Table 1, and

III.2.1 If the examinee identifies a criterion,[a] makes a comparison on the basis of that criterion, and cites a general principle, *then proceed to the next item.*

III.2.2 If the examinee identifies a criterion, makes a comparison on the basis of that criterion, but cites no general principle, then probe: "So (state the criterion mentioned) makes the difference?" with emphasis on the statement of the criterion. Wait for the examinee to respond and *then proceed to the next item.*

III.2.3 If the examinee identifies a criterion, but makes no comparison on the basis of that criterion, then probe: "Could you tell me more about the difference (state the criterion mentioned) makes to your thinking?" Wait for the examinee to respond and *then proceed to the next item.*

III.2.4 If the examinee does not identify a criterion, then probe: "Could you explain a little more what makes you believe one report more than the other?" Wait for the examinee to respond and *then proceed to the next item.*

III.3 If the item tests any principle other than Principle I, and

III.3.1 Same as III.2.1

III.3.2 Same as III.2.2

III.3.3 Same as III.2.3

III.3.4 If the examinee does not identify a criterion, then probe: "Did (state the criterion assumed when designing the item) play any part in your thinking?" If the response is affirmative, then probe: "Could you explain the part it played?" Wait for the examinee to respond and *then proceed to the next item.* If the response is negative, *then proceed to the next item.*

III.3.5 If the examinee explicitly rejects as a criterion the one assumed when designing the item, then probe: "Could you tell me some more about why (state the criterion assumed when designing the item) does not make a difference to your thinking?" Wait for the examinee to respond and *then proceed to the next item.*

III.4 Interviewing strategies for Stage Three:

III.4.1 If the examinee identifies more than one criterion including the one assumed when designing the test, then probe about the assumed criterion.

III.4.2 If the examinee identifies more than one criterion not including the one assumed when designing the test, then probe for the first one identified.

III.4.3 If there is doubt about categorizing a response, then choose the less leading probe, that is, the one that comes first on the list.

---

[a]A criterion is a fact or combination of facts cited in the text that can be used to make a comparison between the observers, observation conditions, or observation reports. In Item 3, the criterion is the combination of facts that Wang was involved in the accident and Vernon was not involved.

The Alberta Journal of Educational Research *Vol. XXXVIII, No. 3, September 1992, 177-188*

Judith K. Bernhard
Ryerson Polytechnical Institute

Gender-Related Attitudes and the Development of Computer Skills: A Preschool Intervention

Computer instructional programs designed to achieve equitable educational outcomes for girls have yielded mixed results. The present intervention involved 59 preschool boys and girls. Within three classrooms, separate girls' and boys' microcomputer groups were formed and matched respectively with girls' and boys' control groups (nature groups). The microcomputer groups received training in open-ended software and were taught basic skills in the LOGO language, over a six-week period. The control groups participated for the same amount of time in a variety of moderately novel, nature-related activities that were designed as gender-neutral. Both groups had female instructors and same-sex pairing. There were two major findings. The first result was that the girls' microcomputer group did not show more positive gender-role attitudes toward computers on the posttest when compared with the girls' control group. Both groups of girls viewed computers as more appropriate for their own gender than on the pretest. The second result was that in spite of equivalent instruction designed to hold the interest of girls, boys completed a significantly greater number of LOGO tasks than girls after the intervention. The question of possible initial advantages of the boys is discussed. Educational interventions with equitable goals face great difficulties in achieving and demonstrating effectiveness. This study suggests that measures of performance should be included in such demonstrations.

Les programmes d'enseignement par ordinateur conçus pour rendre équitables les résultats scolaires des filles ont connu des succès mitigés. La présente intervention a porté sur 59 garçons et filles d'âge préscolaire. Au sein de trois classes, des groupes-microprocesseurs de filles et de garçons étaient formés et appareillés, respectivement, avec des groupes de contrôle de filles et de garçons (groupes-nature). Les groupes-microprocesseurs ont reçu une formation en logiciel "ouvert" (open-ended) et ont appris les rudiments du langage LOGO, pendant six semaines. Les groupes de contrôle ont participé pendant la même durée à une variété d'activités un peu nouvelles, reliées à la nature, et conçues comme sexuellement neutres. Tous les deux groupes avaient des femmes comme instucteurs, et étaient appareillés par sexe. Il y avait deux résultats majeurs. Le premier résultat était que le groupe-microprocesseur de filles n'a pas montré des attitudes plus positives concernant le rôle sexuel à l'égard des ordinateurs au post-test, comparé au groupe de contrôle de filles. Tous les deux groupes de filles ont trouvé l'ordinateur plus approprié pour leur propre sexe qu'au pré-test. Le deuxième résultat était que, en dépit d'une instruction équivalente, faite pour maintenir l'attention des filles, les garçons ont terminé un nombre plus grand, d'une manière significative, de travaux en LOGO que les filles, après l'intervention. Les interventions éducatives avec des buts équitables connaissent de grandes difficultés dans l'atteinte et la démonstration de leur efficacité. Cette étude suggère que des mesures de performance doivent être incluses dans ces démonstrations.

Judith Bernhard is a faculty member at the School of Early Childhood Education. Her research interests include interventions to improve the learning processes and outcomes of girls and ethnoculturally diverse groups.

In fields such as science, mathematics, and computing there are strong patterns of occupational segregation whose effects include lower income, job prestige, and social status for women. The underrepresentation of women in high-level science and computing careers has been the focus of various government efforts to study the inequity and address its causes (e.g., Science Council of Canada, 1982; Status of Women Canada, 1989). Women are inadequately prepared and served by the educational system; they are less likely than men to graduate from university with computer science as a major. In 1988, 21.5% of Canadian bachelor's degrees in computer science were earned by women, as opposed to 78.5% by men (Statistics Canada, 1989). Similarly, women are underrepresented in high school mathematics and computer courses (Ontario Ministry of Education, 1988). These women are said to have "freedom of choice" and "equal opportunity," and clearly they make choices. Yet there is a pattern of avoidance determined in turn by lack of preparation at even earlier levels. Already in elementary school, gender differences in computer attitudes and performance have been consistently noted (Carmichael, Burnett, & Higginson, 1985; Hawkins, 1985). There are, in fact, gender differences in computer use and gender-related attitudes toward computers at the preschool level (Collis & Ollila, 1985; Elliott, 1989; Klinzing, 1985). In particular, girls may view use of the computer as inappropriate to their gender. This study addresses the question: What are effective instructional means of changing these attitudes and improving girls' computer skills?

A number of social factors have been cited as contributing to girls' poor performance in the computer domain. Some of these are beyond the scope of this paper: lack of confidence in computer abilities (Ogletree & Williams, 1990; Wood, Collis, Kass, Kieren, & Therrien, 1988), and girls' beliefs that technological careers are not in their future (Eccles, 1987). Other factors include teacher attitudes (Woodill, Anderson, & Bernhard, 1985), advice of school counselors (Stewart, Hutchinson, Hemingway, & Bessai, 1989), parental encouragement (Carmichael et al., 1985), the nature of the software (Malone, 1981; Turkle, 1984), group dynamics in mixed-sex groups (Guntermann & Tovar, 1987; Webb, 1984), and the ratio of children to computers (Lipinski, Nida, Shade, & Watson, 1986).

Girls' attitudes toward computers have been the focus of several interventions. Two approaches have emerged: those based on providing equal opportunities and those based on special programs for girls to ensure equitable treatment and promote equal outcomes. Collis and Ollila (1990) used an equal opportunity approach in which first graders were exposed to language arts programs. In an earlier study, Collis (1985) had found that by junior secondary school school computer use is seen as a masculine activity (see also Lockheed, 1985; Sanders, 1984). Collis and Ollila (1990) believed that the implementation of a writing program with computers would lead the girls to see that computer use was appropriate to their gender. Despite the intervention, the girls, like the boys, still saw the computer as masculine. The researchers concluded that "gender-related differences in computer use are likely to continue for students as they move through school" (p. 19). Although resistance to change is a significant phenomenon, this conclusion seems premature and unduly pessimistic; it is important to consider possible shortcomings of this unsuccessful

intervention. It is known, for instance, that boys often tend to monopolize computer terminals (Carmichael et al., 1985; Elliott, 1989; Klinzing, 1985). To achieve effectiveness, it seems important to ensure that both sexes have equal exposure time to the computer and to computer instruction.

Recently, an increasing number of interventions based on the equity approach have been published; those cited below used the LOGO computer language. Siann, MacLeod, Glissov, and Durndell (1990) developed a 12-week LOGO program for nine-year-olds. The children worked in pairs with female instructors; half of the pairs were same-sex. By assigned group, children used LOGO for either list processing or turtle-graphics activities. There was no control group with noncomputer activities. Data indicated that gender role differences in attitudes toward computers were significantly reduced regardless of pair composition or type of LOGO used. The reduction of differences was mainly due to the boys becoming much less confident in endorsing such statements as "boys are better at computers." It was noted, however, that the boys tended to "physically crowd girls out" (p. 189). Following the interventions, girls showed only a very small positive increase in response to the statement "I would be good at using computers." Further, the girls generally reported more anxiety after the computer experience. Neither boys nor girls were assessed in computer skills. Based on the disappointing results, the authors recommended that "very careful steps" (p. 190) be taken so as to avoid exacerbating gender differences in computer use. They recommended interventions which "proceed on many fronts" (p. 190); for instance, school-wide programs promoting equity (see Sanders & Stone 1986), formation of girls' computer committees, open-house programs stressing equity, and saturation of an institution with computers so as to eliminate the struggle for access.

Hawkins (1985) and her colleagues conducted a two-year intervention with 8- to 12-year olds. There were no noncomputer control groups. Having noted gender differences in the first year of their study, they attempted to make conditions more favorable to girls. The teachers adopted several new approaches and activities. Projects were designed that were thought to be of special interest to the girls; for example, programming word games. Generally, the teachers tried to spend more time with the girls. Nevertheless, by the end of the second year, gender differences (in favor of the boys) in interest and usage continued to be evident, particularly among the 11- and 12-year-olds. At all levels boys outperformed girls with respect to knowledge of programming.

There have been other LOGO interventions with preschoolers, not specifically designed to be girl-friendly, but which did address gender differences in skill acquisition. Hughes, Macleod, and Potts (1985) intervened with 17 six-year-olds over a five-month period. Children were instructed in mixed pairs. Postintervention assessment of LOGO competence indicated that there was no significant difference between the sexes. Sprigle and Schaefer's (1984) results for LOGO instruction for three- and four-year-olds were similar.

There are strong indications that gender stereotyping for computers peaks in grade 1 (Wilder, Mackie, & Cooper, 1985) and that once these attitudes and patterns are established, they become extremely resistant to change (Collis & Ollila, 1990). Hence it is clearly desirable to carry out interventions in the earliest grades before stereotypic behavior patterns are firmly set. At the time

of this writing, however, no published equity-based intervention studies with preschool children have been found. Although the above studies showed limited results of ensured equal use, it was decided that further investigation of the equal use approach was warranted. It was further decided for purposes of rigor, to use control groups engaged in noncomputer activities. The microcomputer intervention of the present study was designed especially to target preschool girls and had various special features: (a) female role models, (b) same-sex pairs, (c) software judged to be gender-neutral, and (d) ensured, equal use time. The focus of the microcomputer intervention was a controlled assessment of children's learning of the LOGO computer language.

Because there was some evidence that attitudes can be changed, it was predicted that the treatment group's gender-role stereotypic attitudes would decrease significantly in relation to those of the control group. Because the evidence has been mixed and inconclusive as to performance, the following null hypothesis was proposed: boys and girls will perform equally on LOGO skills at the posttest.

Method

Subjects

Two laboratory preschools and a metropolitan daycare center were involved. The subjects were 26 girls and 33 boys (total n=59) between the ages of 3.0 and 6.25 years. Although all subjects had some computer exposure because of the presence of a microcomputer in each classroom, they had no previous experience with the software used in the study. Due to the verbal nature of the measures, the children's receptive vocabulary level was assessed. In order to test whether the groups differed in vocabulary level, the Peabody Picture Vocabulary Test-Revised (PPVT-R, Dunn & Dunn, 1981) was administered. The boys obtained a mean score of 112.5 and the girls a mean score of 111.7. A 2x2 ANOVA with Sex (2) and Group (2) as independent variables and children's PPVT-R scores as the dependent variable revealed no significant differences between the sexes or between the treatment groups. Further, chi-square analysis of a parent questionnaire (Center for Young Children, 1983) revealed no significant differences between the sexes or the groups on computer experience outside the school, or the number of families that owned microcomputers.

Design

The study was a between-within design with three, between-subject factors (Age, Sex and Group) and one, within-subject factor (Occasion; pre- and posttest). To measure the effect of age as an independent variable, a median split by age was used to form two age groups, the first ranging from 36 months to 50 months (3.0-4.12 years; n=29), and the second ranging from 51 months to 75 months (4.25-6.25 years; n=30). Two experimental groups, a boys-only (n=17) and a girls-only (n=12) treatment group ("microcomputer groups") were formed and received instruction in microcomputers. It was not possible for any students to practice their LOGO skills in the classroom, but only during the instructional session. The two groups were matched respectively with a boys-only (n=16) and a girls-only (n=14) control group ("nature

groups") who were instructed in nature-related activities. In all cases, children were taught in same-sex pairs.

Assistants

All measures were administered by two female research assistants unaware of the experimental hypotheses and group assignment. The instructional programs were carried out by two other female research assistants (the "instructors"), also unaware of the experimental hypotheses. All research assistants had completed an undergraduate course relating to computers and young children.

Instructional Material

Microcomputer. The six-hour microcomputer package was based on EZ-LOGO and included the following open-ended software: The Friendly Computer (MECC), Sesame Street Mix and Match (Sesame Street), Facemaker (Spinnaker), Drawing: Early Games for Young Children (Springboard). The package contained a graded series of LOGO tasks beginning with the commands that are used to move the turtle forward or to its right according to its current location. The next set of activities involved exploring the length and width of the screen, directing the turtle over an acetate maze, and "teaching" the turtle to land on a sticker. The last series of tasks included various games:

1. getting the turtle to its box;
2. "teaching" the turtle to travel to each of six stickers pasted on the computer screen;
3. drawing mazes on an acetate transparency, and then "teaching" the turtle to travel through the path. The package was designed so as to permit individual instruction according to each child's progress.

Nature. These activities were designed to be either neutral or in accord with traditional classroom gender roles. Activities involved active exploration, manipulation and discovery by the child, for example, painting with ice cubes, making seed collages, and learning parts of a plant. Later activities included sorting living and nonliving things, dissolving substances, and learning about pitch by blowing on straws of different lengths.

Measures

Activity Object Instrument (AOI). Gender role stereotyping toward computer-related objects and activities was measured with the Activity Object Instrument (Collis & Ollila, 1985), an adaptation of Downing and Thompson's (1977) Activity-Object Opinion Survey, which has been validated with an international population (Downing et al., 1979). The AOI is a booklet of 16 sketches, 8 of objects without a stick figure present, and 8 of stick-figures participating in an activity. Four domains are represented in the 16 items included in the AOI: Four of the sketches are reading-related, four writing-related, four microcomputer-related, and a fourth category includes miscellaneous items. Subjects are presented with a sketch of an object or activity and required to decide if a boy or a girl would be more suitable in the picture. Each domain is scored as follows: A score of 2.00 means that "boy" and "girl" were chosen equally. A score greater than 2.00 indicates that "boy" was selected more often than "girl," and a score below 2.00 that "girl" was selected more often than "boy." The measure for each domain has a range of 0-4. Validity

and reliability of the AOI are reported to be adequate (Collis & Ollila, 1990). In the present study, only scores for the microcomputer domain were used.

LOGO skills tasks. The children's computer skills were measured with the LOGO Skills Tasks (LST), an informal, individually administered series of three tasks developed for the present study. Tasks were to be completed by each child without time limits. Each of the three tasks began with the cursor in the middle of the screen. An initial practice task to ensure comprehension involved asking the child to take the turtle to several places on the screen. The first task consisted of making the turtle go from the center of the screen to the upper, right-hand corner of the screen ("Turtle to sticker"). The second comprised joining the stars at the four corners of the screen ("Joining four stickers"). The third and final task involved leading the turtle through a maze ("Maze"). Each child was assigned a score of 1 or 0 on each task according to whether they were successful or not; the efficiency of their path was not considered. For an overall measure of computer skill, the three task scores were summed. Therefore, this measure had a range of 0 to 3.

Peabody Picture Vocabulary Test—Revised. In order to control for the effects of receptive vocabulary, the Peabody Picture Vocabulary Test—Revised (PPVT-R; Dunn & Dunn, 1981) was administered.

Parent Questionnaire (PQ). The parent questionnaire developed by the Center for Young Children (1983) was designed to elicit information regarding children's exposure to, and use of, electronic media outside the school setting. There were also questions regarding parental views and expectations of children's use of computer technologies. For example, one set of questions asked, "Outside the school does your child have experience with microcomputers, video games, and/or electronic aids?" Another set of questions asked whether there was a computer in the home, who used it, and how often. A third set of questions enquired into parental views of advantages and disadvantages of young children using computers at school. The data from the parent questionnaire were tabulated and analyzed by performing a chi-square analysis in order to determine significant differences between groups or genders.

Procedure

The pretests (AOI, PPVT-R, PQ) were administered three weeks before treatment, and the posttests (AOI, LST) were administered within the week following the termination of the intervention. All test measures were administered by two female research assistants. Tests were individually administered.

Treatment. The microcomputer and nature instructional programs were both carried out over a six-week period; both programs comprised 18 sessions, each of which lasted approximately 15-20 minutes. A female research assistant took the children to a quiet room in the school and worked with them in same-sex pairs.

Microcomputer. The open-ended software was presented during the first two weeks in order to familiarize the children with the names of parts of a computer and the proper handling and loading of disks. After being introduced to basic LOGO commands, children were taught the strategy of pretending to be a turtle as a means of clarifying direction commands. In later sessions involving turtle activities, the children were encouraged to explore

and were verbally reinforced for their efforts. Every effort was made to ensure the child's interest in and understanding of the tasks.

Nature. The instructor presented the materials and suggested the activities that could be carried out with the paint, ice cubes and paper. As the children proceeded, the experimenter made suggestions or comments designed to elicit a variety of exploratory activities. Every attempt was made to get the children interested and engaged in the activities.

Results

Gender Role Attitudes Toward Computers

Children's pretest and posttest scores on the computer domain of the AOI are shown in Table 1. The score indicates the number of times (out of four) that "boy" was selected as more suitable for a presented sketch. A score of 2.00 represents a neutral response, whereas a score less than 2.00 indicates that "girl" was chosen more times than was "boy." Neither the boys' nor the girls' microcomputer group compared with their respective control changed in their gender role attitudes; that is, their view of whether the computer was appropriate to their gender. A significant main effect for Sex was found [$F(1,51)=106.6$, $p<.001$]. Because the boys' responses (mean=3.31) were further from the neutral of 2.00 than the girls' responses (mean=1.33), the boys showed more gender stereotypic attitudes toward computers than did the girls. There was also a significant Sex-by-Occasion interaction [$F(1,51)=10.24$, $p<.002$] indicating that both boys' and girls' tendency to choose own-sex responses increased across the two occasions. No age differences were found.

Computer (LOGO) Skills

Table 2 shows mean scores for the LOGO Skills Tasks (LST). The boys completed a mean of 2.47 tasks (SD=.62) and the girls completed a mean of 1.58 tasks (SD=1.0). In order to determine if this difference was significant, a Mann-Whitney Rank Sum was computed and found to be significant [$U=49.5$, $p<.02$]. Thus, in spite of equivalent instruction, the boys completed a significantly greater number of tasks than did girls.

In order to obtain more detailed information on the pattern of results across the different computer tasks, a chi-square analysis test (with Yates correction) was performed (see Table 2). The girls consistently completed fewer tasks than

Table 1
Means and Standard Deviations for Gender-Typing of Computer Use

Subjects	*Pretest*	*Posttest*
Boys	3.06 (0.99)	3.55 (0.83)
Girls	1.62 (0.85)	1.04 (1.08)
Microcomputer	2.40 (1.16)	2.50 (1.59)
Nature	2.45 (1.21)	2.38 (1.57)
Young	2.44 (0.99)	2.53 (1.58)
Old	2.39 (0.99)	2.30 (1.58)

Table 2
Chi Square Analysis on LOGO Skills Tasks (LST) by Gender

LOGO Task	*Males*	*Females*	*Chi-Square*
1. Turtle to sticker	17/17	9/12	2.43
2. Joining four stickers	15/17	5/12	5.12*
3. Maze	10/17	5/12	0.28

*significant at $p<.05$

the boys, but only one of the three tasks considered alone showed a significant difference, the "Joining four stickers" task).

In order to determine if there were significant age differences in ability to complete the LST, a Mann-Whitney Rank Sum was computed, and it was found that the older children completed a significantly greater number of tasks than the younger children [$U=35.5$, $p<.002$]. A follow-up chi-square analysis (with Yates correction) was applied in order to consider the tasks separately. Table 3 shows the pattern of age-related task performance: the older children completed significantly more tasks than the younger children.

Qualitative Observations

The superior performance of the boys was paralleled by their greater eagerness to learn LOGO. There were several obvious differences between the behavior of boys and girls during the intervention. Girls generally showed a desire to finish the task satisfactorily and sought immediately to return to the classroom, although they had apparently liked the tasks. During teaching they awaited instructions and asked permission to do things. Boys, on the other hand, seemed to find the tasks more interesting and explored eagerly. They raised questions about procedural matters. They shared among themselves to a greater degree than did the girls. Boys often tended to remain at the computer after their tasks were completed. In sum, the boys acted more avidly and inquisitively than did the girls in this computer domain. This difference may have existed in noncomputer domains as well, but no data were gathered that could bear on this issue.

Table 3
Chi Square Analysis on LOGO Skills Tasks (LST) by Age

LOGO Task	*Old*	*Young*	*Chi-Square*
1. Turtle to sticker	14/15	12/14	0.004
2. Joining four stickers	13/15	7/14	2.99
3. Maze	12/15	3/14	7.74*

*significant at $p<.05$

Discussion

Issues Related to Attitudes

Girls. At the posttest, the girls' microcomputer group did not differ from their control regarding perception of the appropriateness of computers for girls. Both groups of girls, regardless of treatment, began by viewing computers as gender appropriate, and both moved in the desired direction of perceiving computers as even more gender appropriate (i.e., "for girls"). The content, that is, subject matter, of the two programs in the present study did not appear to be a major factor in producing changes in the girls' attitudes. It may therefore be argued that, in the absence of maturational changes, the factors common to the two programs may have influenced the girls' attitudes in both groups. (It would seem unlikely that significant maturational changes occurred in the six-week period.) Before discussing what the common elements might be, the present findings are related to those already reviewed. Siann et al. (1990), with female instructors and other girl-friendly measures, reported a reduction of gender differences in attitudes. In view of the present findings, it is entirely possible that the causes of the shift were not connected with the computer-related aspects of the intervention. In attempting to ascertain the effective noncomputer elements of Siann et al., we return to the issue of common elements of the two treatments in the present study: there was an "investigative curriculum"; it was presented in conditions favorable to girls (same-sex pairs, female role models). These are the most obvious elements that may have helped girls in both groups feel more positive about their role in computing. It is also possible that the girls in the microcomputer group talked to the other girls and influenced them with respect to computer-related attitudes. This possibility is somewhat diminished because the microcomputer girls could not practice newly learned skills at the regular classroom computers used by other girls.

In view of the lack of specific results, the direction of Siann et al.'s (1990) comments seems particularly appropriate. They did not suggest abandoning programs, but rather instituting more comprehensive approaches. Approaches with many fronts appear to be necessary to effect changes in educational outcome. Such approaches might include such measures as unequal amounts of time in the girls' favor, special compensatory instruction, social use of computers, parent workshops, and classroom visits by women in traditionally masculine occupations.

Boys. Siann et al. (1990) found that the boys' group after intervention were less likely to agree with statements about boys' superior computer abilities. Hence the question arose for the present intervention, Would the boys' microcomputer group become less stereotyped? Would this group as compared with its control come to believe that the computer was more appropriate to girls (i.e., less exclusively "for boys")? In fact, the boys' microcomputer group showed no such difference compared with their control group. At the posttest, the boys of both groups had undergone rapid changes and believed *more strongly* than at the pretest that computers were for their own gender. Thus, the boys' attitudes were quite sensitive to the experimental manipulation. Of course, all the boys in the present study were in same-sex pairs. Nevertheless, apart from the factor of pairing, the general question remains

unanswered: Are boys, when confronted with an instructional program involving females' increased presence, direction, and modeling likely to become less or more gender stereotyped in their attitudes toward computers?

Issues Related to Skill Acquisition

According to hypothesis B, it was predicted that the boys and girls would perform equally on LOGO-related tasks after the microcomputer instruction. The hypothesis was not accepted. On the LOGO Skills Tasks (LST) administered to the microcomputer group it was found that there were sex differences: boys completed significantly more tasks than girls.

With the benefit of hindsight, one can infer that the features in the computer instructional program intended to benefit the girls were probably insufficient to overcome possible initial advantages. The nature and source of boys' advantages in computer programs as well as math and science programs is extensively documented (see e.g., Harding, 1983, Sanders & Stone, 1986). Advantages may come from (a) initial skills, (b) stronger interest and motivation, (c) gender differences in play styles, and (d) gender differences in sharing and collaborative work.

Boys may have had superior initial skills because of the situation of the computer in their families. Although the data showed no significant gender differences regarding the presence of microcomputers in the home, this did not rule out all gender differences in the amount and manner of usage of microcomputers in the home. Elliott (1989), for instance, found that although 20% of all children had a computer in the home, 70% of boys had used it, compared with only 30% of girls. Furthermore, the boys in Elliott's study had much more knowledge of computer functioning than girls and were more familiar with the terminology relating to computers.

Regardless of the specific source of boys' advantages in the present study, their superior performance is consistent with the findings of Hawkins (1985). The present qualitative observations generally support one of Hawkins' suggestions as to the cause of girls' lower achievement. She labeled this factor *interest*, and it is clearly the same factor that has been here described as eagerness, enthusiasm, and avidity.

Successful performance is a powerful motivator for the development of positive attitudes and expectations. Where performance is lower, an individual may gradually lose motivation, especially where lower performance is consistent with gender-role stereotypes. It seems clear then, that interventions should not merely be aimed at changing gender-role attitudes but also at improving performance. The effect of the present intervention on skills but not attitudes in the microcomputer groups suggests the importance of performance measures in future, equity-based interventions.

Implications for Education

The postintervention gender differences in performance found at this early age indicate that similar conditions of instruction, even girl-friendly approaches, may not be sufficient to change girls' stereotyped perceptions and to impart skills comparable to those of boys. Equitable programs (i.e., those sufficient to achieve equality of outcomes) must specifically address the disadvantages of girls that already exist at the preschool level and do not simply

appear at puberty. These disadvantages need to be dealt with long before the beginning of high school. More comprehensive equity programs are needed to bring girls into the fields of mathematics, science, and computing so that they consider these domains as entirely appropriate to their gender and vital to their careers.

Acknowledgments

This research could not have been carried out without the cooperation of the children and teachers of the Early Learning, Hydrokids, and George Brown centers. Preparation of this article was supported in part by a grant from the Ryerson Polytechnical Institute Faculty of Community Services Research and Scholarship Committee and a doctoral fellowship from the Social Sciences and Humanities Research Council of Canada. Apple Canada graciously supplied the microcomputers. Linda Siegel, Peter Lindsay, Paula Caplan, Sherry Turkle, and Betty Collis provided valuable suggestions. The editorial assistance of Hal White is gratefully acknowledged. Special thanks to Nancy Hutchinson and Robert Short for their support in revision for publication.

References

Carmichael, H., Burnett, D., & Higginson , W. (1985). *Computers children and classrooms: A multisite evaluation of the creative uses of microcomputers by elementary school children.* Kingston, ON: Ontario Ministry of Education.

Center for Young Children. (1983). *Parent computer questionnaire.* Unpublished questionnaire, College Park, MD: University of Maryland.

Collis, B. (1985). Sex-related differences in attitudes toward computers: Implications for counselors. *The School Counselor, 33*(2), 120-130.

Collis, B., & Ollila, L. (1985). *Gender stereotypes about reading, writing, and computer use in preschool, kindergarten, and grade 1 boys and girls.* Unpublished manuscript, University of Victoria.

Collis, B., & Ollila, L. (1990). The effect of computer use on grade 1 children's gender-stereotypes about reading, writing, and computer use. *Journal of Research and Development in Education, 24*(1), 14-20.

Downing, J., Dwyer, C.A., Feitelson, D., Jansen, M., Demppainen, R., Matihaldi, H., Reggi, D.R., Sakamoto, T., Taylor, H., Thackray, D.V., & Thomson, D. (1979). A cross-national survey of cultural expectations and sex-role standards in reading. *Journal of Research in Reading, 2*(1), 8-23.

Downing, J., & Thomson, D. (1977). Sex-role stereotypes in learning to read. *Research in the Teaching of English, 11*, 149-155.

Dunn, I.M., & Dunn, L.M. (1981). *Peabody picture vocabulary test—Revised.* American Guidance Survey.

Eccles, J.S. (1987). Gender roles and women's achievement-related decisions. *Psychology of Women Quarterly, 11*, 135-172.

Elliott, A. (1989). Girls, computers, and early childhood education. In N.J. Kyle (Ed.), *Women as educators in 19th and 20th century Australia.* Wollongong: University of Wollongong.

Guntermann, E., & Tovar, M. (1987). Collaborative problem-solving with LOGO: Effects of group size and group composition. *Journal of Educational Computing Research, 3*(3), 313-334.

Harding, J. (1983). *Switched off: The science education of girls.* York, UK: Longman for Schools Council.

Hawkins, J. (1985). Computers and girls: Rethinking the issues. *Sex Roles, 13*(3/4), 165-180.

Hughes, M., Macleod, H., & Potts, C. (1985). Using LOGO with infant school children. *Educational Psychology, 5*(3-4), 287-301.

Klinzing, D.G. (1985). *A study of the behavior of children in a classroom equipped with computers.* (ERIC Document Reproduction Service No. ED 255 320)

Lipinski, J.M., Nida, R.E., Shade, D.D., & Watson, J.A. (1986). The effects of microcomputers on young children: An examination of free-play choices, sex differences, and social interactions. *Journal of Educational Computing Research, 2*(2), 147-166.

Lockheed, M.E. (1985). Women, girls, and computers: A first look at the evidence. *Sex Roles, 13*(3/4), 115-122.

Malone, T. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 4*, 333-369.

Ogletree, S.M., & Williams, S.W. (1990). Sex and sex typing effects on computer attitudes and aptitude. *Sex Roles, 23*(11/12), 703-712.

Ontario Ministry of Education (1988). *September school report forms,* Curriculum System, Table 11.47P, unpublished data.

Sanders, J.S. (1984). The computer: Male, female or androgynous? *The Computing Teacher, 1*(18), 31-34.

Sanders, J.S., & Stone, A. (1986). *The neuter computer: Computers for boys and girls.* New York: Neil-Schuman.

Science Council of Canada (1982). Who turns the wheel? *Proceedings of a workshop on the science education of women in Canada.* Ottawa: Author.

Siann, G., Macleod, H., Glissov, P., & Durndell, A. (1990). The effect of computer use on gender differences in attitudes to computers. *Computers in Education, 14*(2), 183-191.

Sprigle, J.E., & Schaefer, L. (1984). Age, gender, and spatial knowledge influences on preschoolers' computer programming ability. *Early Child Development and Care, 14,* 243-250.

Statistics Canada. (1989). *Science and technology economic analysis.* Policy sector, industry, science, and technology, unpublished data.

Status of Women Canada. (1989). *Participation of girls and women in math, science and technology.* Paper submitted to the joint meeting of federal/provincial/territorial ministers responsible for the status of women and provincial/territorial ministers responsible for education.

Stewart, H.M., Hutchinson, N.L., Hemingway, P., & Bessai, F. (1989). Program advice given to Canadian adolescents: Effects of student gender, ethnicity, and achievement. *Alberta Journal of Educational Research, 35*(3), 255-266.

Turkle, S. (1984). Women and computer programming. *Technology Review,* November/December, 48-50.

Webb, N. (1984). Sex differences in interaction and achievement in cooperative small groups. *Journal of Educational Psychology, 76*(1), 33-44.

Wilder, G., Mackie, D., & Cooper, J. (1985). Gender and computers: Two surveys of computer-related attitudes. *Sex Roles, 13*(3/4), 215-228.

Wood, P.H., Collis, B., Kass, H., Kieren, T., & Therrien, D. (1988). *Gender-related access patterns and types of computer usage among Canadian adolescents.* Paper presented at the Canadian Society for Studies in Education, Windsor, Canada.

Woodill, G., Anderson, K, & Bernhard, J. (1985). *Sexism, computers, and young children.* Paper presented at the joint meeting of the Association of Educational Data Systems and the Educational Computing Organization of Ontario, Toronto.

The Alberta Journal of Educational Research *Vol. XXXVIII, No. 3, September 1992, 189-204*

Ellen M. Regan
OISE Department of Applied Psychology

and

John A. Ross
OISE Trent Valley Centre

Professional Development for School Consultants: An Evaluation[1]

Most empirical studies reporting effects of consultant training reveal such flaws as inadequately described treatments, use of vague self-report measures, and no comparison/control group. This article describes the effects of a consultant inservice that was focused on specific consulting strategies derived from research on consultant effectiveness and delivered through a program designed according to a "reflective" consultant role orientation. The inservice was assessed through pre- and posttreatment interviews that revealed strategies used by consultants when working with individuals and groups. In the first year of the study, the performance of 18 treatment consultants was compared with 18 matched controls. In the second year the program was revised and delivered to 35 consultants. In the first year the treatment group modestly outperformed the controls; the second year treatment consultants learned substantially more than both groups. The impact of the program was attributed to the use of professional case studies as stimulus for reflection. Consultants are professionals who help other professionals achieve job related objectives. Educational consultants help teachers and other staff implement school change and improve practice. For example, in one Canadian province the function of consultants is to "assist teachers in maintaining proper standards and improving instruction" (Ontario, 1984). Recent events such as reductions in the length of consultants' tenure and in the number of funded positions suggest substantial dissatisfaction with performance. Attempts to improve the quality of educational consultation have tended to focus on selection and supervision strategies, with disappointing results (e.g., Louis & Kell, 1981). The purpose of the research reported here was to develop, implement, and evaluate an inservice approach to enhancing consultant effectiveness.

La plupart des études empiriques indiquant des effets de la formation à la consultation révèlent des défauts tels que des traitements insuffisamment décrits, l'emploi de rapports vagues, et l'absence d'un groupe de comparaison/groupe de contrôle. Cet article décrit les effets d'une formation continue à la consultation axée sur des stratégies spécifiques de consultation s'inspirant de la recherche sur l'efficacité à la consultation, et délivrée à travers un programme fait à partir d'une orientation "réflexive" du rôle du consultant. La formation était évaluée à partir d'entrevues avant et après le traitement dans lesquelles étaient mentionnées les stratégies utilisées par les consultants travaillant avec des individus et des groupes. Pendant la première année de l'étude, la performance de 18 consultants dans un groupe de traitement était comparée à celle de 18 consultants dans un groupe de contrôle. Pendant la deuxième année, le programme était révisé et délivré à 35 consultants. Durant la

---

Ellen Regan is a professor in the Department of Applied Psychology. Her research interests include practitioner beliefs related to professional roles and the professional development of teachers and consultants.
John Ross is a professor in the Department of Curriculum and Head of the Trent Valley Centre. His research interests include instructional improvement and professional development.

première année, la performance du groupe de traitement était légèrement supérieure à celle du groupe de contrôle; pendant la deuxième année, la performance des consultants du groupe de traitement a dépassé de loin la performance des deux groupes. L'effet du programme était attribué à l'emploi d'études de cas professionnelles comme objet de réflexion.

Few studies of consultant inservice (or preservice) have been reported in the literature. Most articles merely provide a rationale for offering consultant training (e.g., Padfield, 1981) or prescribe its characteristics without assessing effects (e.g., Froehle, 1978). The few empirical studies that have been reported are flawed. For example, Broskowski, Khajavi and Mehryar (1973) claimed positive effects for a 12-week training program but the nature of the treatment was only briefly sketched; the effects were assessed by vague self-report measures and there was no comparison group. Goodwin and Coates (1973) provided lectures on theory with simulations and guided practice in lab and school contexts. They found large changes in consultants' behavior, which were associated with effects on teachers and students, but again there was no control group.

Inservice Approach

Inservice Goals

In designing the inservice, primary attention was given to identifying the competencies to be acquired. Although most studies of consultation have reported positive findings (see, for example, the review of 29 school consultation studies in Medway, 1979), only a few investigators have attempted to identify specific consultant behaviors that have an impact. Most inquiries examined the effects on teachers of variations in the *quantity* of consultation. For example, two studies (Loucks, 1982 and Louis, 1981) found that schools where support was provided by two groups of consultants (internal and external) showed greater evidence of change than schools where only one type was involved. Several other studies (Curtis & Watson, 1980; Miles, Fullan, & Taylor, 1978; Runckell & Bell, 1976; Tyler & Fine, 1974) found that the duration of consultant-teacher contact had a positive impact on teacher growth. Other investigators made finer distinctions in terms of the time variable. Loadman and Mahan (1972) found that effective consultants spent more time on inservice and modeling activities, especially demonstration teaching. Louis and Sieber (1979) found that *when* the consultant time was delivered (during implementation rather than up front) mattered more than the total amount of time given.

Few attempts have been made to identify differences in the effects of strategies used by consultants. Firestone and Corbett (1981) were not able to identify any consultant activities that consistently produced change; similar findings were reported by Louis and Kell (1981). Miles, Saxl, and Lieberman (1988) attempted to distinguish between the strategies used by external consultants who were rated as outstanding by project managers and those who were rated average: the differences were small and inconsistent.

In contrast, Ross and Regan (1990) found substantial differences in the strategies used by an exemplary group of experienced consultants compared with those used by inexperienced consultants. They conducted a series of semistructured interviews with 12 curriculum consultants (half inexperienced,

half experienced) from two large school boards, six senior administrators from the same boards, and six additional consultants from a third board with significantly different characteristics. Categories of analysis evolved from the data and were organized in a four-phase model of consultation: initiation, planning, delivery, and follow up. Although similarities in beliefs about the consulting process were found, consultant groups differed strikingly in their methods for helping teachers. These differences were interpreted in the expert-novice framework that has been used for other educational roles (e.g., Bereiter & Scardamalia, 1986; Berliner, 1986; Leithwood & Stager, 1987). The delivery and evaluation strategies used by relative experts suggested a reflective approach to consultation not demonstrated by the novices. The strategies of more experienced "reflective" consultants revealed a problem solving approach in responding to consultee requests and in consultant initiated activity. Less experienced consultants tended to accept consultee definitions of problems and to respond to such requests with pat answers or solutions.

The findings from the Ross and Regan (1990) study were used to construct a profile of consultant development containing 11 dimensions with a hierarchy of strategies for each. The four levels in each dimension represented incremental steps toward effectiveness. Ten of these dimensions and the anchoring points (upper and lower levels) on each were derived directly from the expert-novice study data. The intermediary levels were derived from the same source or were interpolated by the researchers. In all instances, the levels were validated by reviewing the school change literature.

For example, for Profile Dimension 3: *Determining the needs to be met*, the expert-novice data revealed that inexperienced consultants tended to impose their own personal agendas (Level 1) in planning activities to meet needs. One inexperienced consultant reported, for example, that his workshops replicated a provincial inservice program with specific prescribed objectives, that is, "it's quite a program ... very prescriptive ... very tight." Other inexperienced consultants described objectives for their consulting activities as deriving from personal agendas or their general knowledge of teachers in their boards. The researchers added to these data the observation that novice consultants frequently impose school system priorities without regard for the perceived needs of teachers (Level 2). In contrast, experienced consultants were concerned with responding to teacher expectations. For example, one experienced consultant referred to the importance of obtaining accurate data on teacher needs (Level 3 strategy). This consultant referred specifically to a situation in which she talked with both the teacher and principal "beforehand ... that way everybody is brought on board ... and you can help the teacher a little bit more because you have more information." Experienced consultants talked about determining needs as a process of negotiating expressed needs and priorities (Level 4). For example, one experienced consultant observed, "there's a (system) plan ... you try to get them to plug into that plan." Another experienced consultant talked about using the system plan and "data we had collected from people in their brainstorming sessions."

School change research concerned with the identification of adult learner needs provides support for the nature and sequence of all four levels on this dimension. In the literature on techniques for identifying needs of adult

learners (e.g., Coffing & Hutchison, 1974; Griffin, 1978), there is agreement that appropriate procedures consist of identifying a desired state, collecting accurate information about current status, and measuring discrepancies between the two. The three elements of this shared procedure are increasingly prominent in higher levels of the consultant profile. Imposition of the consultant's agenda (Level 1) involves none of these elements. Imposition of system priorities (Level 2) involves one element, specification of desired state. Response to expressed needs (Level 3) involves two elements—actual status and perceived discrepancies. The (Level 4) strategy, negotiation of expressed needs and system priorities, involves all three elements.

The evidence used in constructing the four strategy levels for the first 10 profile dimensions is reviewed in Regan and Ross (1991). The 11th profile dimension, *reflecting within the role*, was added later as a summative orientation to the role. It resulted from observing how consultants responded to inservice activities. This dimension distinguished between consultants who appear to be unreflective about their work and those who consciously analyze activities and outcomes on a regular basis with a view to self-improvement. The resulting growth continuum (shown in Figure 1 as the Consultant Profile) describes a gap between actual and ideal performance that was to be overcome through the inservice.

Inservice Design

The consultant inservice was founded on three principles discussed and supported in the literature as guides for helping professionals improve their practice. The principles were:

1. focus on specific strategies and skills because training that combines knowledge acquisition with heavy emphasis on skill development is potentially more effective (Goodwin & Coates, 1973; Griffin & Barnes, 1986; Miles et al., 1988);
2. provide opportunities to reflect on one's own practice (Clark, 1988; Schön, 1987);
3. provide opportunities to practice skills in framing problems and in experimenting with these frames in actual practice environments (Schön, 1983, 1987).

We saw the strategies of expert consultants as actions that could be taught to less experienced consultants by providing tasks that linked the findings from research to the consultant's own experience. The tasks that were designed collectively approximated the Gliessman, Grillo, and Archer (1989) stages of problem solving cited by Kagan (1990) in her discussion of research related to teacher self-reflection. Consultants were engaged in analyzing problems, generating problem solutions, interpreting solutions in the light of additional information, and considering the consequences of different courses of action. Kagan's review found that verbalizing one's reflections in writing or sharing with another is important to the process of self-reflection. Accordingly, the inservice tasks and materials provided for recording individual and group reflection as well as the opportunity to share reflections and problem solving strategies through group discussion.

In accordance with these principles, the inservice was designed to engage participants in specific tasks within a continuing cycle of reflection-planning-

Dimension	Level 1	Level 2	Level 3	Level 4
Before consultation begins				
1. Identifying the recipients of consultation	individual teacher	teacher in context of other teachers)	teachers and administrative team in the school	teachers and administrators in adjacent schools
2. Recruiting the consultation team	consultant works alone	consultant teams with one other agent (e.g., another consultant)	consultant teams with two other agents (e.g., consultants and expert teachers)	consultant teams with three other agents (e.g., expert teachers, school administrators, and other consultants)
Planning activities to meet needs				
3. Determining the needs to be met	imposition of consultant's personal agenda	imposition of system priorities	(a) response to expressed needs of teachers based on limited data (b) response to expressed needs of teachers identified systematic data search	negotiation of expressed needs and system priorities
4. Individualizing within the consultation	treat all teachers the same	plan extra follow-up for enthusiastic teachers	provide options for all teachers to select from	assign each teacher to most appropriate option
5. Providing practice within consultation sessions	no planned practice	teachers discuss teaching	(a) teachers experience activities as students (b) teachers simulate teaching (e.g., each other or plan units)	teachers teach children
6. Providing practice between consultation sessions	no planned practice	practice with no planned feedback	practice with feedback	provide support that enables teachers to design self-directed practice and self-appraisal
7. Using in-school expertise	no planned teacher sharing	sharing encouraged but not scheduled	planned sharing by expert teachers to spread ideas to others	planned sharing by all teachers to consolidate new ideas
Evaluation and follow-up				
8. Response to teacher feedback during delivery	ignores feedback	expands or contracts treatment of objective	shows how consultation meets teacher needs	shifts to a different system objective that better meets teacher needs
9. Identifying types of data to assess consultation	perceived satisfaction of the teacher	perceived usefulness to the teacher	objective indicators of teacher change	objective indicators of student change
10. Using evaluation data	no use	planning subsequent sessions with the same teacher	planning subsequent sessions with similar teachers	refinement of consultation strategies
11. Reflecting within the role	no reflection	reflection on incidents	reflection on specific instances within a framework	systematic experimentation to develop new strategies

References to the *teacher* in the profile should be read as the *consultee.*

Figure 1. Consultant profile.

practice-reflection. The target skills in these tasks were the specific strategies identified in the Consultant Profile. Eight case studies developed by the research team provided the anchor for the inservice. The cases were organized in four sets of two. Each set described a problem commonly experienced by consultants; one case presented the problem as experienced in a one-to-one consultation and the other presented the same problem in a group setting. Both cases in a given set addressed the same profile dimensions and reflected a Level 1 strategy for each dimension. Collectively the eight cases addressed the first 10 dimensions of the Consultant Profile. Each session was structured around Case Study Analysis, Action Planning, and, after on-site implementation of the plan, Feedback.

In Case Study Analysis there was a brief period of consultant personal reflection followed by a longer period of group sharing and analysis. Working in groups of three to four, consultants identified the strategies described in the pair of cases and their practice implications. Information on the strategies of effective consultants, based on earlier research and illustrated in the Profile, was then provided by the research team. The group leader's role was to clarify Profile and task dimensions as needed, keep the discussion on task, encourage sharing by all group members, and help the group achieve closure regarding its task. Members of the research team served as team leaders in the first year of implementation; they were replaced in the second year by consultants.

The Action Planning tasks involved analysis of a self-identified "personal case in the past" and development of a future action plan based on perceived discrepancies between current practice and the higher levels of the Consultant Profile. In addition to describing strategies to be implemented in a future consultation, consultants elaborated their plans by supplying details of how they would proceed, what they saw as alternative courses of action, and how they planned to assess the effects of their new strategies. The group leader's role in support of these activities included clarifying the task and providing feedback on the action plans, for example, encouraging precision and specificity in the plans and suggesting alternatives to be considered. Between sessions action plans were implemented by consultants in their everyday roles over a five- to seven-week period.

The Feedback tasks provided consultants with the opportunity to share and critique the implementation of their action plans. Each group was responsible for identifying the important practice implications of the strategies they used. The group leader's role in this activity was to ensure that all group members shared/critiqued their experiences and were able to identify their own growth.

Implementation and Evaluation of the Professional Development Program

The program was piloted in a truncated form in two school boards. Fourteen consultants in one board and 18 in the other analyzed one pair of case studies, reflected on their current practice, designed a plan to improve that practice, and received feedback on the implementation of their plans. Anecdotal evidence provided by participants suggested positive outcomes, but no systematic data were collected. The full program was implemented and evaluated with two different groups of consultants over a two-year period (1988-1989, 1989-1990).

The specific questions to be investigated in the research reported here were:

1. Did the inservice have a beneficial impact on the strategies used by consultants?
2. Did revisions to the inservice program in the second year increase its impact?

Study Year 1 (1988-1989)

Sample/Procedure

Eight consultants in one large board in central Ontario and 10 consultants in a medium-sized board located 125 km from the first volunteered to be the treatment sample. Consultants were given release time to attend sessions in their respective board offices. The treatment sample was matched in terms of experience and subject area responsibility with 18 volunteer consultants from adjacent boards.

In September 1988 individual interviews were conducted with subjects in the two samples. Treatment sessions (three full days and two half days) began in late October and continued until late April 1989. Post interviews for both treatment and control groups were conducted in May 1989. The Interview Guide asked subjects to describe two of their recent consultations, one in which they were working with individuals and the other with groups. Subsequent questions probed the specific strategies consultants used. The interviews were audiotaped and verbatim transcripts were made. The sample of consultant strategies described were coded as levels of the Consultant Profile. Each transcript was independently coded by two trained research assistants; disagreements between coders were resolved through discussion.

Univariate statistics for pre and post scores were generated. Because pretest interviews correlated with post scores, standardized residuals were produced by regressing post over pre scores. Differences between treatment and control groups were determined by conducting one-way analyses of variance on the residuals.[2]

Results

A substantial amount of data was missing on three Profile dimensions. "Reflection within the role" had not been included in the original version of the interview guide because our appreciation of this variable did not emerge until near the end of the inservice. Attempts to recapture performance on this dimension by examining responses to other items on the pretest administration failed. Because a pretest score could be confidently determined for only a few members of the sample, the item was dropped from the analysis. There was also a problem with "using assessment data." On the pretest several respondents in both samples misunderstood the question; the problem was resolved on the post administration by adding further probes. Finally, several consultants experienced difficulty responding to the dimension concerned with "teacher feedback during delivery of consultation." Some persisted in describing actions that would be taken at a subsequent consultation (rather than during the session in which the problem arose); others indicated they were unable to answer because they had never encountered negative feedback during delivery. These responses, almost half the total for the teacher feedback items in both administrations, were coded as missing.

Table 1 displays the means and standard deviations of each profile dimension for each group on each test occasion; the table also shows the average based on 10 dimensions. (The average was used as the summative outcome measure rather than the total because of missing data.) Table 1 indicates that the means increased slightly over the duration of the project in both samples. The analyses of variance in the residuals found that the treatment group significantly outperformed the control group overall [$F(1,34) = 4.22$, $p<.048$], although the differences on the individual dimensions reached significance only for determining needs [$F(1,34) = 5.35$, $p<.027$] and response to teacher feedback [$F(1,8) = 7.57$, $p<.025$]. The overall effect size of the program using the formula of Glass, McGraw, and Smith (1981) was a modest .26.

Discussion

Consultants enrolled in the inservice were effusive in their appraisals of the program. They perceived that it had a large effect on their practice, far more than was indicated by the measured outcome. The most likely explanation for the participants' enthusiasm is that the program filled a void: there were no competing programs that the consultants could compare it with and it provided an otherwise rare opportunity for consultants to talk to peers about their work. But why did the inservice have such a small effect on what consultants do? Several explanations came to mind.

1. The upper levels of the Profile were developed by selecting the best efforts of an outstanding group of consultants, and the definition of these levels was further enhanced by incorporating findings from research on school change. The difficulty in moving from one level of the Profile to another may have been underestimated. In several of the sessions the inservice designers attempted to address several different components of a single dimension or combined several complex strategies. Some participants may have experienced overload, particularly since the inservice did not set priorities among the dimensions and did not attempt to provide different levels of inservice corresponding to the specific needs of each of the candidates. In addition, there may have been insufficient practice to consolidate the new levels of performance into existing practice. Each dimension was addressed only once; no further opportunity for practice was given, and there was no attempt to accumulate the dimensions in successive action plans.
2. The academics who delivered the inservice may have been too active in the deliberations of the small groups. Observations by research team members during the inservice sessions found that comments/questions tended to be addressed to the research team members rather than to fellow consultants, thus influencing the nature of group interaction and the topics of discussion.
3. Responses to certain tasks and materials suggested a need for revision to highlight particular strategies, clarify task requirements, and modify procedures. For example, there was confusion about the meaning of some dimensions, specifically reacting to teacher feedback during delivery of consultation and individualization within consultation. Because attention to these consulting dimensions is so dependent on consultant on-the-spot awareness and sensitivity, an alternate means of treating these dimensions

Table 1
Means and Standard Deviations for 1989 Study Outcome Measures

Profile Dimension	*Pretest*						*Posttest*					
	Treatment			*Control*			*Treatment*			*Control*		
	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD	N	$\bar{X}$	SD
Identifying consultation recipients	18	1.69	0.30	18	1.58	0.19	18	1.86	0.51	18	1.61	0.21
Recruiting consultation team	17	1.79	0.69	18	1.39	0.63	17	2.24	0.73	18	1.61	0.61
Determining needs	18	2.69	0.42	18	2.61	0.45	18	3.07	0.48	18	2.65	0.56
Individualizing within consultation	18	1.72	0.96	16	1.81	1.17	18	1.94	1.16	17	1.76	1.03
Practice within sessions	16	2.56	1.00	17	2.31	0.98	17	2.60	0.78	18	2.49	0.90
Practice between sessions	13	2.04	1.01	15	1.90	0.63	16	2.13	0.63	17	2.32	0.66
Using in-school expertise	16	2.03	0.64	17	1.94	0.56	16	1.75	0.68	18	1.89	0.63
Response teacher feedback	10	2.20	0.79	9	2.11	0.78	12	2.08	0.51	9	1.56	0.73
Identifying assessment data	17	2.59	0.92	18	2.42	1.02	18	2.69	0.96	18	2.17	1.03
Using assessment data	12	2.63	0.48	9	1.94	0.81	17	2.56	0.50	18	2.36	0.64
Average score	18	2.16	0.40	18	1.99	0.32	18	2.32	0.31	18	2.07	0.30

Scores potentially range 1-4 on each item

needed to be developed. In addition there was confusion about purposes that could be addressed with a goal setting activity directed at developing a framework for team and individual consulting.

Study Year 2 (1989-1990)

Sample/Procedure

Ten boards in eastern and central Ontario ranging in size from 6,800 to 52,600 students agreed to participate by sending a team consisting of one experienced consultant who would act as team leader and two to four inexperienced consultants. There were 35 volunteers in the program; all were viewed as competent by their supervisors. The inservice program was delivered in rotating sites with different boards serving as hosts for each of the five sessions (three full days, two half days).

The consultant sample was relatively young: most were aged 36-40 and only four were over 50. Two thirds were female and half the total sample had interrupted their careers, in most cases for family reasons. Most had not been in their current role for very long: only three had more than three years experience in their current position and almost half had none. When previous consulting experience was added to current tenure, the mean was 2.66 years as a consultant. Only two had consultancy experience outside their own board. A majority had responsibility for helping both elementary and secondary teachers; the remainder worked solely in the elementary panel, with more in the primary than in the other divisions. Most were very experienced teachers (17 years on average); five had served as principal or vice-principal prior to becoming a consultant. Most had served on various board curriculum committees although a fifth had not. Formal qualifications were biased toward professional rather than academic training. Virtually all had obtained at least one subject specialist certificate and most reported additional practical qualifications. Five had not completed an undergraduate degree and only three had acquired a master's. One had finished a doctorate.

The Year 2 sessions, delivered from late October 1989 to late April 1990, were similar to those in Study Year 1 with refinements in materials and procedures in response to feedback from Year 1 participants. A main goal of the revisions was to provide better focus. For example, in Year 1 both of the selected dimensions for a given pair of case studies were highlighted in the questions following each case description; in Year 2 the questions for one case highlighted one dimension, and the second case highlighted the other dimension. Another way of achieving better focus was by addressing fewer aspects of a profile dimension in a single session, providing participants with a clearer sense of inservice parameters, and designing new activities to help consultants set personal learning goals. The increased focus made it possible to provide additional practice activities for each dimension without extending the time allotted to each session.

Each board team had a formally designated team leader who was assigned the task of guiding the team through the inservice activities and stimulating reflection within the group. The researchers observed the team's deliberations but did not intervene except to respond to explicit requests for procedural direction. Separate sessions with team leaders were used to review progress, to plan subsequent activities, and to provide additional information. With

respect to the latter, researcher meetings with team leaders were used to introduce and discuss the two Profile dimensions that Study Year 1 data indicated were poorly grasped by the participants. These dimensions concerned with individualizing and responding to feedback, respectively, within consultation, were discussed in the context of helping team members develop an awareness of when on-the-spot adjustments are demanded by the consulting situation.

Pre and post interview data were collected[3] from the participating consultants using the same instrument as in Year 1. The pretest interviews were conducted in September 1989; the post interviews were administered in May 1990. All interviews were audiotaped. Two trained research assistants coded the tapes, assigning Consultant Profile levels to each subject on the basis of the examples of individual and group consulting that were provided. Any disagreements between coders were resolved through discussion. In addition, each consultant was asked to nominate a teacher familiar with his or her work. The teacher was asked to rate the performance of that consultant on each of the dimensions of the profile, using a five point scale that included "don't know." These teacher ratings were summed and correlated with the sum of the profile placements from the interview.

When all interviews were transcribed and coded, pre and post means and standard deviations were generated. T-tests were used to assess the significance of pre/post changes. The data from Year 2 were combined with the data from Year 1 in an analysis of covariance where the dependent variable was post interview score, the independent variable was treatment condition (Year 1 control, Year 1 treatment, and Year 2 treatment) and pretest score was the covariate.

In June a confidential report was sent to each participant describing the inservice team's perception of the extent to which the individual's strategies had changed during the program. Specific evidence for these perceptions was provided in the form of excerpts from the pre and post interviews. Subjects were invited to add missing data and/or correct interpretation of this data; no additional feedback was received.

Results

There were a large number of uninterpretable reactions to the interview questions on "response to teacher feedback during delivery." Given the number of cases for which one or both scores was missing, the item was deleted. The remaining 10 items were summed to produce a total score. All items correlated positively with the sum of the remaining items; the alpha reliability on pretest scores (n=28) was .58; on the post (n=33) it was .79. The sum of the items also correlated positively with the sum of the teacher ratings (r=.67) but the correlation was based on only 15 cases because of missing data.

Table 2 displays the means and standard deviations for each dimension and the total score. The table shows that Consultant Profile scores increased from pre to post; the differences were significant overall and on eight of the 10 dimensions.

Comparisons were also made between the scores of the Year 2 treatment condition and the Year 1 treatment and control groups. The dependent variable was the posttest average of nine profile interview items that were com-

Table 2
Means and Standard Deviations for 1990 Study Outcome Measures

Profile Dimension	Pretest N	Pretest $\bar{X}$	Pretest SD	Posttest N	Posttest $\bar{X}$	Posttest SD
Identifying consultation recipients	35	1.69	0.53	35	2.17	0.84***
Recruiting consultation team	35	1.96	0.85	35	2.47	0.74**
Determining needs	35	2.73	0.81	35	2.83	0.79
Individualizing within consultation	32	1.63	0.99	33	2.20	1.19**
Practice within sessions	33	2.46	0.93	35	3.20	0.69***
Practice between sessions	33	1.95	0.71	35	2.39	0.69**
Using in-school expertise	31	2.00	0.82	35	2.34	0.85
Identifying assessment data	35	1.89	0.83	35	2.54	0.93***
Using assessment data	32	2.28	0.70	35	2.76	0.65**
Reflecting within the role	35	1.83	0.38	35	2.03	0.45*
Total	28	20.91	3.56	33	24.97	4.77***

Scores potentially range 1-4 on each item and 4-40 overall.
*$p<.05$. **$p<.01$. ***$p<.001$.

mon to the three data sets. (The "reflecting within the role" dimension was not included because it was missing from Year 1 data; "responding to teacher feedback during delivery" was omitted because it was missing from Year 2 data.) The analysis of covariance showed that posttest scores were strongly influenced by pretest performance [$F(1,68)=23.25$, $p<.001$] and to a lesser degree by treatment condition [$F(2,68)=5.35$, $p<.007$]. The Newman-Keuls procedure for the multiple range test indicated that the Year 2 treatment group learned significantly more than the Year 1 control and the Year 1 treatment. The revision of the materials and procedure used to implement the treatment in the second administration appeared to have a positive effect. In this analysis the differences between the Year 1 treatment and Year 1 control did not reach significance, in contrast with the small but significant finding reported in Table 2, because the dependent measure was based on nine rather than 10 items.

As in the Year 1 study, observations made during the inservice sessions were used to refine the program further. For example, it was observed that labels (e.g., "the empty request", "the hostile audience") that had been used as case study titles should be deleted because they were a distraction to some participants.[4]

Discussion

The primary purpose of the Year 2 study was to find out if revisions to the inservice materials and procedures had a positive effect on performance. Comparison with the 1989 treatment and control groups indicated that the 1990 treatment was superior. The revision of inservice materials and procedures had a positive effect, most likely by increasing the reflectiveness of the participants while working through the professional development tasks. This

argument, along with audiotape evidence on consultant interactions during the inservice, is expanded in Ross and Regan (1991b).

General Discussion

The findings from this research will be of primary interest to those concerned with the enhancement of the consultant role. The findings suggest that it is possible to improve the quality of consultants' strategies through a collaborative program of input, reflection, and action based on the inservice principles described above. But the findings will also be of interest to those who are experimenting with the use of case studies to stimulate professional growth. Several collections of teacher case studies have been recently published (Greenwood & Parkay, 1989; Kowalski, Weaver, & Henson, 1990; Shulman & Colbert, 1987, 1988; Silverman, Welty, & Lyon, 1991) for use in preservice and inservice programs. There have been few evaluations of the case study approach to teacher development, although some evidence of positive impact has been provided by Welty, Silverman, and Lyon (1991), and to a lesser extent by Shulman and Colbert (1989). The research reported here corroborates the claim that reading and discussing the experiences of other professionals can make a positive contribution to professional growth. The findings, particularly the differences in the outcomes of the 1989 and 1990 versions of the program, also suggest that certain conditions play a critical role in the effective use of case studies.

1. The problems elaborated in cases must be so familiar to inservice participants that they are able to spontaneously place themselves in the setting. Without such familiarity the connections between one's own experience and that of others as a foundation for new learning is unlikely to be made. In this program the cases were generated from reports of consultants working in similar settings to those of the inservice participants, and specific discussion probes were provided to help them make connections between the case, their past actions, and future intentions.
2. Each case should be more than a report of an event: it should represent an instance of a larger class and illustrate a particular principle of practice (Shulman, 1986). McAninch (1991) criticized existing collections of teacher case studies for failing to make explicit linkages between case content and theories of what teachers should know and do. In the consultant inservice the cases were written to focus attention on specific dimensions of consultant strategies that distinguished relative experts from novices.
3. Learners need to be motivated by the case discussion to change their practice. In the 1989 study some groups questioned whether strategies described in the higher levels of the Profile were superior to their current methods. In the 1990 version the discussion probes were revised so that groups did a card sort in which they rank ordered alternative ways of dealing with the problem outlined in the case. When the Profile dimensions were distributed, consultants invariably found that their rank ordering of the approaches conformed to the sequence in the Profile. This activity enhanced the credibility of the Profile, displaced their commitment to existing routines, and motivated change in professional practice.
4. The case studies should provide an achievable challenge. The cases used in the consultant inservice were not worst case scenarios. Rather, they sug-

gested how in a given situation an inexperienced but conscientious consultant without a solid repertoire of strategies might respond to a significant problem. Consultants had to make use of their own experience and professional knowledge, as well as data from the research literature, to come up with better solutions for case study problems.

5. The length of the case studies may influence inservice learning. At times participants felt not enough information was presented in the cases and they rationalized the behavior of the consultant described in the case by suggesting that there may have been unknown factors preventing the protagonist from acting differently. At other times case study details cued discussion of some aspect of consulting irrelevant to the task at hand (e.g., discussion of attitude/behavior of a particular principal). The case studies used in both years of the consultant inservice were considerably shorter (about 200 words prior to discussion probes) than most teacher case studies that have been reported.

Our findings, especially in Year 2 of the study, lead us to conclude that the systematic treatment of new learnings initiated in the case studies in each session, and carefully elaborated in the planning, practice, and subsequent reflections on practice, offer some explanation of the success of the inservice treatment. Certainly other elements might affect achievement such as the duration of the program, the training of group leaders, the goals and objectives of sessions, grouping strategies, and, as already suggested, procedures for feedback and reflection. The relative contribution of each factor in the model to the final outcome was not addressed in the research reported here and further studies systematically varying each design element would be appropriate.

Notes

1. Funding for this research was provided by grants from the Ontario Ministry of Education to the Ontario Institute for Studies in Education. The views expressed in the article do not necessarily reflect the views of the Ministry. The authors thank Anne Hogaboam-Gray who contributed to the data collection and Peter Gray who contributed to the data analysis. The authors would also like to express their appreciation to the consultants who participated in the research. A version of the article was presented at the annual meeting of the Canadian Association for Curriculum Studies, Kingston, Ontario in June 1991.
2. Interview data were supplemented with additional information that are not reported here. These included (a) records of action plans and reports of implementation experiences generated during inservice activities; (b) field notes produced by researchers during the sessions; (c) job descriptions, workshop outlines and the like provided by consultants; and (d) audiotapes of deliberations among consultants when engaged in case study analysis, action planning and feedback tasks.
3. Data were also collected on antecedent conditions that affected consultant learning during the inservice. These are reported in Ross and Regan (1991a).
4. Other differences between Year 1 and Year 2 implementations of the program, are reported in Ross and Regan (1991b).

References

Bereiter, C., & Scardamalia, M. (1986). Educational relevance of the study of expertise. *Interchange, 17*(2), 10-19.

Berliner, D. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*(7), 5-13.

Broskowski, A., Khajavi, F., & Mehryar, A. (1973). An evaluation of an inservice training seminar on community consultation and education. *Journal of Community Psychology, 1*(2), 174-176.

Clark, C. (1988). Asking the right questions about teacher preparation: Contributions of research on teacher thinking. *Educational Researcher, 17*(2), 5-12.

Coffing, R., & Hutchison, T. (1974). *Needs assessment methodology: A prescriptive set of rules and procedures for identifying, defining, and measuring needs.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Curtis, M., & Watson, K. (1980). Changes in consultee problem clarification skills following consultation. *Journal of School Psychology, 18,* 210-221.

Firestone, W., & Corbett, H. (1981). Schools versus linking agents as contributors to the change process. *Educational Evaluation and Policy Analysis, 3*(2), 5-17.

Froehle, T. (1978). Systematic training for consultants through competency-based education. *Personnel and Guidance Journal, 56,* 436-441.

Glass, G., McGraw, B., & Smith, M. (1981). *Meta-analysis in social research.* Beverley Hills, CA: Sage.

Gliessman, D.H., Grillo, D.M., & Archer, A.C. (1989, March). *Changes in teacher problem solving: Two studies.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Goodwin, D., & Coates, T. (1973). Increasing teacher effectiveness through social systems change: Training school psychologists as change agents. *California Journal of Educational Research, 25*(3), 147-156.

Greenwood, G., & Parkay, F. (1989). *Case studies for teacher decision making.* New York: Random House.

Griffin, G., & Barnes, S. (1986). Using research findings to change school and classroom practices: Results of an experimental study. *American Educational Research Journal, 23,* 572-586.

Griffin, W. (1978). Educational needs: Definition, assessment and utilization. *School Review, 86,* 382-394.

Kagan, D. M. (1990). Ways of evaluating teacher cognition: Inferences concerning the Goldilocks jungle. *Review of Educational Research, 60,* 419-469.

Kowalski, T., Weaver, R., & Henson, K. (1990). *Case studies on teaching.* New York: Longman.

Leithwood, K., & Stager, M. (1987). *Components of expertise: "Artistry" in principals' problem solving.* Paper presented at the Canadian Society for the Study of Education, Guelph.

Loadman, W., & Mahan, J. (1972). The external consultant and curriculum change strategies. *Theory into Practice, 11*(5), 329-339.

Loucks, S. (1982). External and local facilitators. In S. Loucks (Ed.), *People, policies and practices: Examining the chain of school improvement. Vol. 4: A study of dissemination efforts supporting improvement* (pp. 43-92). Andover, MA: The Network.

Louis, K. (1981). External agents and knowledge utilization: Dimensions for analysis and action. In R. Lehming & M. Kane (Eds.), *Improving schools: Using what we have* (pp. 168-211). Beverly Hills, CA: Sage.

Louis, K., & Kell, D. (1981). *The human factor in dissemination: Field agent roles in the organization context. Linking R&D with schools.* Cambridge, MA: ABT. (ERIC Document Reproduction Service No. ED 207 266)

Louis, K., & Sieber, S. (1979). *Bureaucracy and the dispersed organization: The educational extension experiment.* Norwood, NJ: Ablex.

McAninch, A. (1991). Casebooks for teacher education: The latest fad or lasting contribution. *Journal of Curriculum Studies, 23*(4), 345-355.

Medway, F.J. (1979). How effective is school consultation? A review of recent research. *Journal of School Psychology, 17,* 275-282.

Miles, M., Fullan, M., & Taylor, G. (1978). *Organizational development in schools: The state of the art. Vol. III: OD consultants/OD programs in school districts.* New York: Center for Policy Research.

Miles, M., Saxl, E., & Lieberman, A. (1988). What skills do educational "change agents" need? An empirical view. *Curriculum Inquiry, 18*(2), 157-193.

Ontario, Government of. (1984). *Regulation 262. Revised regulations of Ontario, 1980, as amended to O. Reg. 761/82 under Education Act (Elementary and secondary schools and schools for trainable retarded pupils—general).* Toronto: Queen's Printer.

Padfield, C. (1981). Why school boards should have consultant development programs. *Education Canada, 21*(2), 38-41.

Regan, E., & Ross, J.A. (1991). *Professional development activities for the reflective consultant: Techniques for learning how to help schools.* Toronto: OISE Publications.

Ross, J.A., & Regan, E. (1990). Self-reported strategies of experienced and inexperienced consultants: Exploring differences. *Alberta Journal of Educational Research, 36,* 157-180.

Ross, J.A., & Regan, E. (1991a). *Antecedent conditions affecting learning in an in-service program for consultants.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Ross, J.A., & Regan, E. (1991b). *Sharing professional experience: Talk and achievement in adult cooperative learning groups.* Mimeo. Peterborough, ON: OISE Trent Valley Centre.

Runckell, P., & Bell, W. (1976). Some conditions affecting a school's readiness to profit from OD training. *Education and Urban Society, 8,* 127-144.

Schön, D. (1983). *The reflective practitioner: How professionals think in action.* New York: Basic.

Schön, D. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions.* San Francisco, CA: Jossey-Bass.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Shulman, S., & Colbert, J. (Eds.). (1987). *Mentor teacher casebook.* San Francisco, CA: Far West Laboratory for Educational Research and Development.

Shulman, S., & Colbert, J. (Eds.). (1988). *Intern teacher casebook.* San Francisco, CA: Far West Laboratory for Educational Research and Development.

Shulman, S., & Colbert, J. (1989). Cases as catalysts for cases: Inducing reflection in teacher education. *Action in Teacher Education, 11*(1), 44-52.

Silverman, R., Welty, W., & Lyon, S. (1991). *Case studies for teacher problem solving.* New York: McGraw-Hill.

Tyler, M., & Fine, M. (1974). The effects of limited and intensive school psychologist-teacher consultation. *Journal of School Psychology, 12*(1), 8-16.

Welty, W., Silverman, R., & Lyon, S. (1991). *Student outcomes from teaching from cases.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

The Alberta Journal of Educational Research *Vol. XXXVIII, No. 3, September 1992, 205-218*

Marziya Yasmin

and

Baha Abu-Laban
University of Alberta

Ethnicity and Occupational Inequality: A Reconsideration

There is disagreement in the current literature on social stratification concerning whether ethnicity has an impact on socioeconomic status. This article addresses this issue by introducing a new conceptual-methodological perspective. Using data from the Public Use Sample Tapes of the 1981 Canadian Census, the study employs multiple indicators of ethnicity, incorporates a broad set of social background variables into the research design and examines both inter- and intra-ethnic occupational inequality. The results indicate that social background variables, in particular gender and education, are more important than ethnicity in explaining occupational inequality. However, the results also suggest the continuing relevance of ethnicity in explaining some aspects of inequality. The research and practical implications of these results are discussed.

Il y a un désaccord dans la littérature actuelle de la stratification sociale concernant l'effet de l'appartenance ethnique sur le statut socio-économique. Cet article traite de cette question en introduisant une nouvelle perspective conceptuelle et méthodologique. Se servant de données provenant des Public Use Sample Tapes *du recensement canadien de 1981, l'étude utilise des indicateurs multiples de l'appartenance ethnique, incorpore un large éventail de variables d'ordre social dans le plan expérimental et examine aussi bien l'inégalité inter-ethnique que l'inégalité intra-ethnique dans l'emploi. Les résultats indiquent que les variables d'ordre social, en particulier le sexe et l'éducation, sont plus déterminantes que l'appartenance ethnique dans l'explication de l'inégalité dans l'emploi. Cependant, les résultats suggèrent aussi la pertinence de l'appartenance ethnique dans l'explication de quelques aspects de l'inégalité. La recherche et les implications de ces résultats sont présentées.*

Introduction

The issue of ethnic inequality has been a major research focus in Canada in recent years. Much of the credit for this goes to Porter (1965) whose seminal work has stimulated considerable interest and set the stage for an ongoing debate in this area. Porter maintains that ethnicity has played an important role in class formation in Canada:

> Immigrants from many different cultures have had to fit into a social class structure which, in Canada, is a reflection of the economic position of the many ethnic groups, both Canadian-born and immigrant, who make up the population. In this selecting and sorting out of migrants of different ethnic back-

---

Marziya Yasmin is a research associate in the Department of Sociology.
Baha Abu-Laban is a professor of sociology and Associate Dean (Research and Graduate Students) in the Faculty of Arts.

> grounds into various occupations and so into the class system, a number of factors have operated in varying intensities at different times in Canada's history. (p. 60)

A main theme of Porter's work concerns the persistence over time of a general pattern of vertical arrangement of ethnic groups. According to him:

> Germans, Scandinavians, and Dutch are nearest to the British in their occupational levels ... Italians, Polish, Ukrainians, and groups from southeast Europe are still at the lower end of the occupational spectrum ... The French could be placed somewhere between these last and the groups from northern Europe. (p. 90)

Porter (1965, p. 68) cogently argues that the prevailing racial attitudes in Canada operated in such a way as to funnel the less preferred immigrants (e.g., East Europeans) into a lower occupational or "entrance status"; and to restrict their mobility and that of their descendants. He further argues that in both the industrial and agricultural sectors, ethnic differences have been important in the expansion of the bottom layer of the stratification system. Although there may be local and regional variations in the relationship between ethnic affiliation and social class, the general pattern remains. Apart from everything else, Porter clearly saw rising tension between ascription and achievement as two distinct principles of resource allocation.

The substance of the current debate concerning Porter's thesis is whether ethnic differences do play such an important role in class formation. The works of Richmond (1967a; 1967b), Blishen (1970), Li (1978), Forcese (1980), Richmond and Kalbach (1980) and Lautard and Loree (1984) tend to support Porter's thesis. In contrast, however, the works of Pineo (1976), Darroch (1980) and Reitz (1980) have demonstrated a weak influence of ethnicity on socio-economic status. More recent research evidence reported by Satzewich and Li (1987) shows that income disparity among ethnic groups has tended to persist despite an appreciable reduction in occupational disparity.

A more challenging position has been taken by Li (1988) in his study of ethnic inequality, where he argues that inequality can be better explained by class relations than by ethnicity. Li acknowledges that both class and ethnicity produce inequality in the capitalist system, but

> they tend to operate differently. The source of class inequality.... stems from the unequal relationships to production. Ethnic inequality, however, surfaces in the capitalist labour market. In other words, class-based inequality is fundamental in capitalist production, whereas ethnic and racial divisions are not the direct result of relations of production. (pp. 51-52)

Thus a substantial portion of income disparity may be attributed more to class than to ethnicity (Li, 1988, p. 138).

Although Li's (1988) study presents an interesting articulation of class and ethnicity, it nevertheless raises important questions regarding the conceptual and methodological issues involved in this area of study. These issues draw attention to the need for the development of a new research strategy.

Using a strategy that involves a less limiting research procedure, this article addresses three major areas of concern in the study of ethnicity and occupational inequality in Canada. First, it seeks to reassess Porter's thesis that ethnic

affiliation is a correlative factor in the assignment of occupational roles and thus in the class system. Second, it seeks to determine the extent to which ethnicity, in comparison with other factors, helps to explain occupational disparity. Finally, in order to probe the dynamics of inequality a little more fully, the article seeks to identify correlative factors in occupational inequality within ethnic groups (i.e., intra-ethnic occupational inequality). The overall results from this study may shed some light on whether the above noted contradictions in the literature may be attributed to methodological differences or to a genuine change in class dynamics in Canadian society.

Critical Reflections

Li (1988) correctly notes that "one of the problems surrounding the debate of the vertical mosaic concerns the different ways ethnicity and class are measured in each study" (p. 61). With reference to ethnicity, although it is a compound concept made up of both subjective and objective elements, it has often been defined in terms of only one element, be it subjective or objective. An example of a subjective definition is *ethnic identification*, whereby the individual expresses a positive personal attitude toward a given group. On the other hand, an objective definition involves *classification* by others on the basis of physical (e.g., skin color) and/or sociocultural characteristics (such as language or religion). Moreover, some definitions of ethnicity continue to be characterized by a high degree of ambiguity. For example, in the 1981 Canadian Census, which is a major source of information for social scientists, ethnic origin is determined by the following question: "To which ethnic or cultural group did you or your ancestors belong on first coming to this continent?" The responses to this question are difficult to interpret to say the least, particularly since the respondents were allowed to indicate single or multiple origins, with no instruction given on whether to trace ethnic origin through the father's or mother's side.

The works of some of the earlier researchers such as Porter (1965), Darroch (1980), Lautard and Loree (1984) who followed the Canadian Census definition of ethnic origin in their studies, raise much concern about the validity of conceptualizing ethnic origin. Other researchers like Blishen (1970), Richmond (1967a; 1973) and Boyd (1985a) used country of birth or national origin for examining inequality. In this connection, it has been pointed out by Pineo and Porter (1985) that ethnicity loses much of its meaning among native-born Canadians with subsequent generations, which restricts the use of this variable to the foreign-born members only. Ethnic language was another indicator that some studies employed for examining ethnic inequality as evidenced by the works of Tepperman (1975), Boulet and Veltman (1981) and McRoberts (1985). Again, like place of birth, language is likely to be of significance mainly to foreign-born respondents.[1]

Although each of these indicators is more or less limited in defining ethnicity, it is reasonable to assume that the use of all three indicators, that is, ethnic origin, place of birth, and language simultaneously, may provide a more comprehensive definition of ethnicity. The value of using multiple indicators of ethnicity is shown in the works of Goldlust and Richmond (1974) and Ornstein (1981), even though these studies are restricted to small samples. More recently, Li (1988) employed ethnic origin and nativity for examining

inequality, but the thrust of his analysis appears to be more on ethnic origin than on nativity. The present research employs all three of the indicators mentioned above and includes a large national sample.

With reference to the measurement of class, there are major differences among the earlier studies. Few of these studies have conceptualized class in Marxian terms, that is, in terms of relations of production. For this reason, Li's (1988) study is an important contribution in this regard. Li's analysis of class is based on employment status (as defined in the 1981 Canadian Census). The three categories of employment status included in his study are "self-employed with paid help," "self-employed without paid help" and "employees or workers." The latter category, "employees or workers," has been further subdivided into "managers and supervisors," "professionals" and "workers." Li maintains that employees engaged in managerial or supervisory occupations have some control over money capital and considerable control over physical capital which refers to actual production and labor of others. However, a detailed examination of the managerial category indicates wide occupational differences ranging from production management to post office management. Clearly the degree to which employees in this category exercise control over capital and labor is extremely variable. This is a serious drawback of Li's classification scheme.

Moreover, Li's (1988) attempt to examine the impact of social class on income is tautological in the sense that one of the underlying components in conceptualizing class position is ownership of property, which is often indicated by income.[2]

One of the most widely accepted ways of measuring class position involves the use of such indicators as occupation, income, and education, or often a combination of these variables commonly referred to as socioeconomic status. Although this method is gradational in nature and hence not grounded in a theory of class as a relational concept, the resulting hierarchical categories would most likely reflect significant economic and behavioral differences among the people (Porter 1965, p. 10). The present study follows the gradational mode of class analysis, but for technical reasons it uses occupational rather than socioeconomic status.[3]

Turning now to other issues, a further limitation observed in earlier studies concerns the bivariate analyses of ethnicity and social class. Such an approach tends to overlook the effect of ethnicity on social class while controlling for the effect of additional variables. To overcome this, proponents of the status attainment model (Blau & Duncan, 1967; Cuneo & Curtis, 1975; Duncan & Duncan, 1968; Li, 1978; and Boyd et al., 1981) sought to examine the impact of ethnicity on socioeconomic status by including certain social origin and achievement oriented variables. In this context, a number of researchers have elicited the importance of level of education in explaining ethnic inequality. For example, Li (1978) demonstrates that occupational differences among male immigrants are determined in part by differences in education. Other researchers such as Ornstein (1981) have demonstrated that education plays a more important role than ethnicity in explaining occupational disparity. The present study deals with this aspect by examining first the role of ethnicity in explaining differences in level of education and, second, the role of ethnicity in explaining occupational inequality by controlling the effects of education.

A systematic review of earlier studies also reveals that they have tended to neglect the inclusion of demographic variables. To further our understanding of ethnic inequality, it is important to widen our framework of inquiry and include important demographic variables. Studies utilizing this approach have been sporadic and often confined to a single demographic factor. For example, some researchers have examined the degree to which inequality in socioeconomic status can be explained by gender (Abella, 1984; Boyd, 1985b; 1990; and Fox & Fox, 1986); or by age (McRoberts et al., 1976; and Pineo & Porter, 1985); or by family size (Kenneth, 1970; Pineo, 1985; Trovato & Grindstaff, 1986); or mobility status (Blau & Duncan, 1967; Sandefur & Scott, 1981; and Sharma, 1980). In this study an attempt is made to introduce all these demographic variables along with ethnic variables to examine the net effect of ethnicity on socioeconomic status.

The mode of inquiry followed by researchers of ethnic inequality also warrants some attention. Two main perspectives were followed in this regard. First, some researchers followed an exploratory mode of inquiry and examined the extent to which ethnicity had an influence on social class (Blishen, 1970; Darroch, 1980; Lautard & Loree, 1984; Ornstein, 1981; Porter, 1965). Other researchers approached their investigation by assuming prevalence of disparity in socioeconomic status among ethnic groups and thus proceeded to explain intra-ethnic inequality (Cuneo & Curtis, 1975; Boyd et al., 1981; Boyd, 1985a). The first mode of inquiry can be regarded as an important step toward understanding the role of ethnicity in explaining the dynamics of class inequality, whereas the second approach can be viewed as critical for analyzing class and inequality within various groups. Thus the two approaches can be regarded as complementary; jointly they provide a more comprehensive understanding of the interplay between ethnicity and social class. The present study, therefore, includes both of these research foci.

Admittedly the above noted problems are not exhaustive of the conceptual and methodological issues confronting this research area. However, we are in a position to address them in the context of the present research design. Specifically, we will examine (a) the influence of ethnicity, defined in terms of three indicators, on occupational status; (b) the net influence of ethnicity on occupational status, while incorporating a set of demographic variables in a broader model of ethnic variables, and (c) intra-ethnic occupational inequality.

Research Procedure

Data for this study were obtained from the *Public Use Sample Tapes* of the 1981 Census of Canada, representing two percent of the original sample and including individuals in the age range of 15 to 65 years (N=333,576). This age range represents the work force in the Canadian Census. The variables included in the study are discussed below, but it should be noted that the definitions of all these variables were derived from the Dictionary of the 1981 Census of Canada.

Ethnicity. As indicated earlier, an important feature of this study is that ethnicity is defined in terms of three indicators, namely, ethnic origin, place of birth, and use of home language. While all three of these indicators have been used in the studies reviewed above, the tendency has been to define ethnicity in terms of one of them and not all three together. Given the compound nature

of the concept, we believe that the use of multiple indicators in this instance is methodologically less limiting than the use of only one. The three indicators used in the study are explained briefly below.

Ethnic origin refers to the ethnic group to which the respondent or the respondent's ancestors belonged on first coming to Canada. Six ethnic groupings, primarily based on their broad geographic origin and period of arrival, have been identified. These are the British, French, North-West Europeans, South-East Europeans, non-Europeans, and a final grouping made up of individuals with multiple origins.

Place of birth, also referred to as nativity, is the second indicator of ethnicity. This was entered as a dichotomy of foreign-born and native-born.

The Canadian Census provides information on both mother tongue and home language. Whereas mother tongue refers to the language first learned in childhood and still understood by the respondent, home language refers to the language actually spoken at home by the respondent at the time of the census. For the purpose of this study, home language was regarded as the more valuable indicator of the two. Home language, the third indicator of ethnicity, was entered as a trichotomy, including use of English, use of French, and use of other languages at home.

Occupation. The preceding review of the literature shows that ethnic inequality has been variously studied in terms of class, socioeconomic status, or occupational inequality. As discussed above, the measurement of class from census data seems to present serious problems. Also, although socioeconomic status is a widely accepted measure, there were technical problems in using it in this study[3]. Thus it was concluded that for present purposes occupation was the most useful measure for examining ethnic inequality. This conclusion is not without merit since it is based on the widely held assumption that a person's occupation very often mirrors his or her status or position in society and forms an important criterion for stratification. Moreover, occupation tends to reflect both education and income.

Occupation was entered by using 16 categories of the Canadian Classification and Dictionary of Occupations. These were: Managerial, Natural Sciences, Social Sciences, Teaching, Medicine and Health, Artistic/Literary, Clerical, Sales, Service, Farming, Other Primary Occupations, Processing Occupations, Machining and Repair, Construction Trades, Transport/Equipment, and Other Occupations. After testing these categories in relation to income and education, we collapsed them into three manageable categories representing high, medium, and low occupational status. Of the 16 categories listed above, the first six were combined to define high occupational status; the next three (Clerical, Sales, Service) were used to identify medium occupational status; and the remaining seven categories were combined to identify low occupational status. These occupational status categories were found to resemble those identified by Pineo, Porter and McRoberts (1977, p. 95).

Demographic (social background) variables. Another important feature of this study is the attempt to examine the relationship between ethnicity and occupational status while controlling for the effects of certain demographic or social background variables. The control variables incorporated into the design of the present study are those indicated by our critical review of the literature, namely: age, gender, family size, mobility status, and level of educa-

tion. Leaving age and gender aside, the census definitions of the remaining variables are as follows.

Family size refers to the number of persons in a family unit (as defined by the 1981 Canadian Census). Ten categories of this variable have been entered, ranging from one to 10 members in the family.

Mobility status has been dichotomized between movers and nonmovers, with movers being individuals who did not live in the present residence five years ago.

Level of education refers to the highest level of schooling attained and has been entered as an ordinal variable with 11 categories.

Multiple regression analysis. The procedure of multiple regression was employed for analyzing the data because this procedure allows the researcher to examine the effect of the independent variables on the dependent variable while controlling the effects of other variables in the model. This is a powerful statistical procedure well suited for analyzing the relationship between ethnicity and occupational inequality as it helps to overcome the limitations of the bivariate analyses of ethnic inequality that have been followed in much of the literature.

The analysis is based on standardized (*B*) rather than unstandardized (b) regression coefficients. Despite the relative instability of standardized coefficients across samples, they have been used in the analysis because they provide a more appropriate means of interpreting the effects of variables using different scales. As Pedhazur (1982) observes:

> Unlike the b's, *B*'s are scale-free and therefore can be compared across different variables. It is this property of the *B*'s that appeals to a great many researchers, who use their relative magnitude as an indication of the relative importance of the variables with which they are associated. (p. 247)

To sum up, the research strategy followed in this study is unique and arguably less limiting because it combines the following features: use of a large national sample, use of multiple indicators of ethnicity, incorporation of numerous background variables into the research design, examination of inter- and intra-ethnic occupational inequality, and use of multiple regression.

The Findings

In the light of earlier viewpoints regarding the two modes of inquiry into ethnic inequality, the analysis was carried out in two stages. The first stage treated ethnicity as a determinant of occupation and education and examined the extent to which ethnic and demographic variables explained variation in occupational status and level of education. In the second stage, factors explaining intra-ethnic occupational inequality were examined.

Considering the first stage of analysis, Table 1 shows two different models that were employed for explaining the influence of ethnic and demographic variables on occupational status and level of education respectively.[4] The results of model one are presented in columns 2 and 3. The table shows that this model explains 15% of the variance in occupational status. On the whole, the variables representing ethnicity (i.e., ethnic origin, nativity and home language) appear to be weak in accounting for variation in occupational status, though use of English as a home language appears to have the strongest

Table 1
Occupational and Educational Inequality by Selected Ethnic and Demographic Variables

	Occupation		*Education*	
	b	*B*	b	*B*
Ethnicity Variables				
Ethnic Origin				
British	−.011	−.009+	−.036	−.056*
French	.005	.003+	−.072	−.102*
NW European	.016	.007**	.019	−.017*
SE European	.058	.027*	−.071	−.066*
Multiple Origins	−.032	−.014*	−.013	−.011+
Nativity	.062	.019*	.161	.098*
Home Language				
English	.065	.049*	.118	.174*
Education	.420	.212*	—	—
Demographic Variables				
Gender	.684	.273*	.063	.049*
Age	−.004	−.055*	−.003	−.069*
Family Size	−.060	−.065*	−.032	−.070*
Mobility	.119	.047*	.121	.095*
Constant	.703		1.292	
Multiple R	.392		.204	
R^2	.154		.041	
Adj R^2	.154		.041	
Std. Error	1.150		.620	
DF	12/333563		11/333563	
F	2292.607		592.086	
N	333576		333575	

+Statistically not significant.
*Statistically significant at .01 level.
**Statistically significant at .05 level.

influence (*B.049*) on occupational status. This means that with an increase in the proportion of respondents who use English as home language, occupational status tends to rise.

Among the demographic variables, gender is the most influential variable in explaining variation in occupational status (*B.273*), which in this case means that with an increase in the proportion of male members there is a likely rise in occupational status. Level of education is the next most important variable influencing occupational status (*B.212*), indicating that with an increase in the proportion of those with a high level of educational attainment, there is a concomitant rise in occupational status. The influence of family size and age on occupational status (*B−.065* and *B−.055* respectively) is relatively weak. The negative coefficients of these variables indicate that with an increase in size of families and with age, occupational status tends to decline. Mobility status has the weakest influence on occupational status, with a beta coefficient of .047. It is worth noting, however, that the positive association between mobility and occupational status supports earlier observations by Sharma (1980) and San-

defur and Scott (1981) and possibly reflects the expectation of occupational reward for those who are geographically more mobile.

A somewhat different picture emerges when the influence of ethnic and demographic variables on level of education is examined in model two. It is apparent from column 4 in Table 1 that only 4% of the variation in level of education is explained by this model (compared with 15% of the variation in occupational status explained by model one).

Although the same ethnic variables are used in models one and two, it is clear from Table 1 that these variables have a relatively stronger influence on level of education (model two) than on occupational status (model one, columns 2 and 3). With specific reference to model two, the most noteworthy finding is that use of English as home language has the strongest influence on level of education, with a beta coefficient of .174, which is higher than all the other beta coefficients in the model. This means that with an increase in the proportion of those who use English as home language, there is a likely rise in level of education. The influence of nativity on level of education is relatively weak (*B.098*). With reference to ethnic origin, the French and the South-East European groups show a strong negative association with level of education (*B–.102* and *B–.066* respectively). This finding shows some similarities with earlier observations by Ornstein (1981) to the effect that the French and the Italian ethnic groups had low scores on education.

Among the demographic variables, mobility status appears to be the most influential variable in explaining level of education (*B.095*), which suggests that with an increase in the proportion of members who are geographically mobile, there is a likely rise in level of education. This contrasts with a main finding of the previous model where gender and not geographical mobility was the most influential variable in accounting for variation in occupational status. This finding relating to model one does not necessarily mean that mobility status is of no relevance in explaining occupational inequality. Rather, it may suggest that those individuals (both male and female) who are geographically mobile are more likely to access educational opportunities, which in turn may positively influence their occupational achievement.

Finally, in model two, the influence of family size (*B-.070*) and age (*B-.069*) on educational attainment is relatively weak and that of gender is the weakest (*B.049*). These results probably reflect the democratization and hence greater accessibility of education in Canada today compared with the 1950s and early 1960s.

Turning now to intra-ethnic occupational inequality, Table 2 shows the effects of ethnic and demographic variables on occupational status.[5] Six regression models are presented, one for each ethnic group. The model representing the French ethnic group (column 3) explains 18% of intra-group variation in occupational status. For all the other models, the variance explained ranges from 13 to 14%.

Both the British and the North-West European ethnic groups show some similarity. For example, for both of these groups gender emerges as the most important variable in explaining within-group occupational inequality (*B.271* and *B.296* respectively), followed by level of education (*B.202* and *B.170* respectively). The variables next in order of importance are age, family size, and mobility status. Both use of English as home language and nativity have

relatively weak influence on occupational status. The French and the South-East European ethnic groups also demonstrate that gender has the strongest influence in explaining intra-group occupational inequality (*B.298* and *B.278* respectively), followed by level of education (*B.225, B.172* respectively). But for the French ethnic group, use of French as home language (*B–.074*) and family size (*B–.070*) appear to be the next most important variables in explaining internal differences in occupational status. The negative coefficient associated with use of French as home language indicates that with an increase in the proportion of French-Canadians who use French as home language, there is a likely decline, albeit small, in occupational status. For the South-East European group, nativity (*B.069*) and mobility status (*B.050*) are the next most influential variables in explaining intra-group variation in occupational status.

For the non-European ethnic group as well as those with multiple origins, level of education appears to be the most important variable in accounting for internal variations in occupational status (*B.235* and *B.237* respectively), followed by gender (*B.225* and *B.212* respectively). For the non-European ethnic group, use of English as home language (*B.084*) and nativity (*B.062*) are the next most influential variables affecting occupational status. But for members with multiple origins, family size (*B–.117*) is the most influential variable affecting occupational status, following level of education and gender.

A noteworthy feature that emerges from Table 2 is that for ethnic groups of British origin and those with multiple origins, the influence of home language

Table 2
Occupational Inequality by Nativity, Home Language, Gender, Age, Family Size, and Mobility Status of Six Ethnic Groupings

	British B	*French* B	*NW European* B	*SE European* B	*Non-European* B	*Mult. Origins* B
Ethnicity Variables						
Nativity	–.011	–.033	.004	.069	.062	–.014
Home Language						
English	.013	—	.050	.041	.084	.033
French	—	–.074	.016	–.012*	—	—
Education	.202	.225	.170	.172	.235	.237
Demographic Variables						
Gender	.217	.298	.296	.278	.225	.212
Age	–.090	–.021	–.085	–.041	–.004	–.036
Family Size	–.067	–.070	–.078	–.048	–.056	–.117
Mobility	.045	.062	.031	.050	.014	.035
Constant	.938	.249	1.250	.719	.632	1.128
Multiple R	.375	.426	.384	.362	.371	.361
R^2	.140	.182	.147	.131	.137	.130
Adj R^2	.140	.182	.147	.131	.137	.130
Std. Error	1.129	1.219	1.607	1.098	1.150	1.094
F	3041.729	2926.982	653.442	595.038	619.452	364.465
DF	7/130103	7/91938	8/30165	8/31370	7/27096	7/16984
N	130111	91946	30174	31379	21704	16992

Note: All the coefficients in this table are significant at the .01 level, with the exception of that denoted by an asterisk.

on intra-group occupational inequality is relatively weak. In contrast to this, the influence of home language on occupational status is somewhat stronger among ethnic groups of South-East European and non-European origin. It is interesting that these contrasting results mirror the degree of linguistic homogeneity among the different ethnic groups. For example, over 90% of British-origin Canadians as well as those with multiple origins use English as home language, whereas only 39.2% of South-East Europeans and 37.2% of those of non-European origin use English as home language (1981 Census of Canada). In the Canadian context, the striking differences in terms of home language within the South-East European and non-European ethnic groups are apt to be associated with differences in occupational status.

On the whole, the evidence from this study shows that for all ethnic groups demographic factors rather than ethnic factors are more important in explaining occupational disparity. Also, of all the demographic variables examined, gender and level of education have the strongest influence on occupational inequality. In fact a consistent pattern of strong influence of gender and level of education on occupational status emerges for all the models (except in model two where level of education is the dependent variable). This consistent pattern brings into focus the dynamic role of gender and level of education in the formation of occupational hierarchies in society at large and also across ethnic groups.

Conclusion

While this study reveals some interesting patterns, one would have to be cautious about overgeneralizing. For example, like most of the earlier studies, one of the limitations of the present study lies in using occupation as an indicator of (or substitute for) social class position. The question that remains unanswered is: If class measurement was properly grounded in the Marxian tradition of class theory, would ethnicity still play an important role in determining class inequality?

The inclusion of several demographic variables was an important aspect of this study. The results show that these variables, particularly gender, and also education tend to have a stronger impact on occupational status compared with ethnic variables. Significantly, these results in general are consistent with earlier studies that employed a multivariate rather than a bivariate research design. Judging from these results, we may be justified in concluding that the debate over Porter's thesis is to some degree a debate over method. With particular reference to gender, the evidence underlines the prevalence of a male dominated core labor market and directs attention to the need to investigate more thoroughly the dynamics of gender inequality in the labor market.

It is quite likely that since the publication of Porter's (1965) work the changing labor market conditions and the demand for recruiting immigrants with relatively high educational and occupational qualifications have mitigated the influence of ethnicity on occupational inequality. The start of this trend was evident to Porter (1965) when he declared that one of the reasons for the changes in relative social status "is the difference between ethnic groups in the occupational level of their immigrants" (pp. 86-88). For example, he underlined the exceedingly high representation of professionals (scientists and engineers) among Asian immigrants between 1951 and 1961. In

this context, it is significant to note that Canada's preference for highly qualified immigrants continued through the 1960s and 1970s. It is also quite likely that among native-born Canadians ethnic origin may carry a different meaning today than in the past. Pineo and Porter (1985), for example, in their study of current ethnic identity of Canadians, found that about 86% identified themselves as *Canadians*. These issues, notably the seemingly changing meaning of ethnicity, merit further investigation in the context of a comparative research design.

The fact that ethnic variables in general had a relatively weaker influence on occupational status than demographic variables is not to suggest the irrelevance of ethnicity in Canadian society. For one thing, use of English as home language was found to have a positive influence on level of education, which in turn played a critical role in explaining occupational inequality. Also, income disparity among ethnic groups remains, despite an overall reduction in occupational disparity (see, e.g., Satzewich & Li, 1987; and Rashid, 1985). Indeed, ethnicity may still play an important role in advancement and promotion on the job as well as in the skill requirements of occupations filled by members of different ethnic groups. Relevant to this, Boyd (1985a, p. 401) noted that foreign-born ethnic groups may have difficulties in obtaining recognition for educational credentials in the Canadian labor market and this may limit their life chances. Thus we need to examine the influence of ethnicity on positional advantage within and across occupations.

The implications of the results of this study for educational practice may be stated briefly. Educators are well aware of the tension between the merit and ascriptive criteria in occupational selection and placement and also in the assignment of income. It may be reassuring that the results point to a strong influence of education on occupational status. However, the results also show continued tension between achievement and ascription. The tension between these two principles of allocation is most intense in relation to gender, with women being the more disadvantaged gender group. It also exists, albeit to a much lesser degree, in relation to ethnicity. What this means is that preservice and inservice teacher training programs need to address (or continue to address) gender role and, where necessary, racial biases in the system. As well, school curricula need to be reexamined with a view to correcting these biases. Obviously, educators, school boards, and school teachers and counselors can play a critical role in the move toward gender and ethnic equity.

Similarly, the tension between achievement and ascription is apparent in who gets educated. Educational and hence occupational opportunities are not equally available to all ethnic groups. For example, among the disadvantaged groups are those who are less proficient in the English language. To equalize educational opportunities, school systems need to give higher priority to English as a Second Language (ESL) and other remedial language programs. The emphasis on language proficiency can also be of value to adult immigrants and not only to their children. ESL programs are particularly important because the bulk of immigrants to Canada in the 1990s are projected to be from non-English speaking countries. While such programs may not eliminate ethnic inequalities entirely, they may further weaken the thesis that ethnic affiliation is a correlative factor in the assignment of occupational roles.

Notes

1. Isajiw (1981), on examining language retention of four ethnic groups, has found that retention of ethnic language decreases considerably in the second and subsequent generations.
2. Following Wright's (1977; 1978; 1979) criteria of social class, Li (1988) classifies the "self-employed with paid help" as "employers," and "self-employed without paid help" as "petty bourgeoisie" (p. 68). The rationale for this classification is based on the assumption that both the employers and the petty bourgeoisie own and control money capital and physical capital of production. It appears from this classification that ownership of capital is regarded as a "given" criterion for the employers and the petty bourgeoisie. From this perspective it would be tautological to examine the influence of capital on income.
3. At the initial stage of the analysis socioeconomic status (a composite index of occupation, education, and income) was included but due to multicollinearity between socioeconomic status and education (which is an independent variable included in the analysis), socioeconomic status had to be replaced by occupational status.
4. Of the six ethnic categories, five have been entered as dummy variables with values of 1 in the multiple regression equation. The sixth category, namely, the non-European group, is not included in Table 1 because it has been assigned a value of 0 and treated as a base against which the other categories could be compared. Also, the beta coefficients of French as home language are not included in the table due to multicollinearity between this variable and French ethnic origin on the one hand (.856), and between it and use of English as home language on the other (–.838).
5. Two independent regression analyses were carried out for the British, French, and non-European ethnic groupings, and for those with multiple identity. This was done because of multicollinearity between use of English and use of French as home languages for these four groupings. In the first analysis, use of English as home language was excluded; in the second analysis, use of French as home language was excluded. For the ethnic groupings labeled British, non-European, and multiple identity, use of English as home language was reported because the regression coefficient of this variable was higher than the one associated with use of French as home language. For the French ethnic group use of English as home language was found to be statistically insignificant, and as a result only the analysis that included use of French as home language was reported.

References

Abella, R.S. (1984). *Equality in employment: A royal commission report.* Ottawa: Minister of Supply and Services.

Blau, P.M., & Duncan, O.D. (1967). *The American occupational structure.* New York: Wiley.

Blishen, B.R. (1970). Social class and opportunity in Canada. *Canadian Review of Sociology and Anthropology, 7,* 110-127.

Boulet, J., & Veltman, C. (1981). Socioeconomic achievements of Montreal language groups in 1971. *Canadian Review of Sociology and Anthropology, 18,* 239-248.

Boyd, M. (1985a). Immigration and occupational attainment in Canada. In M. Boyd, J. Goyder, F.E. Jones, H.A. McRoberts, P.C. Pineo, & J. Porter (Eds.), *Ascription and achievement: Studies in mobility and status attainment in Canada* (pp. 393-445). Ottawa: Carleton University Press.

Boyd, M. (1985b). Educational and occupational attainments of native-born Canadian men and women. In M. Boyd, J. Goyder, F.E. Jones, H.A. McRoberts, P.C. Pineo, & J. Porter (Eds.), *Ascription and achievement: Studies in mobility and status attainment in Canada* (pp. 229-296). Ottawa: Carleton University Press.

Boyd, M. (1990). Sex differences in occupational skill: Canada, 1961-1986. *Canadian Review of Sociology and Anthropology, 27,* 285-315.

Boyd, M., Goyder, J., Jones, F.E., McRoberts, H.A., Pineo, P.C., & Porter, J. (1981). Status attainment in Canada: Findings of the Canadian mobility study. *Canadian Review of Sociology and Anthropology, 18,* 657-673.

Cuneo, C., & Curtis, J. (1975). Social ascription in the educational and occupational status attainment of urban Canadians. *Canadian Review of Sociology and Anthropology, 12,* 6-24.

Darroch, A.G. (1980). Another look at ethnicity, stratification and social mobility in Canada. In J.E. Goldstein & R.M. Bienvenue (Eds.), *Ethnicity and ethnic relations in Canada* (pp. 203-230). Toronto: Butterworths.

Duncan, B., & Duncan, O.D. (1968). Minorities and the process of stratification. *American Sociological Review, 33,* 356-364.

Forcese, D. (1980). *The Canadian class structure.* Toronto: McGraw-Hill Ryerson.

Fox, B., & Fox, J. (1986). Women in the labour market 1931-81: Exclusion and competition. *Canadian Review of Sociology and Anthropology, 23*, 1-21.
Goldlust, J. & Richmond, A.H. (1974). A multivariate model of immigrant adaptation. *International Migration Review, 8*, 193-225.
Isajiw, W.W. (1981). *Ethnic identity retention*. Research Paper No. 125, Centre for Urban and Community Studies, University of Toronto.
Kenneth, C.W. (1970). *An introduction to population*. Chicago: Dorsey Press.
Lautard, E.H. & Loree, D.J. (1984). Ethnic stratification in Canada, 1931-1971. *Canadian Journal of Sociology, 9*, 333-344.
Li, P.S. (1978). The stratification of ethnic immigrants: The case of Toronto. *Canadian Review of Sociology and Anthropology, 15*, 30-40.
Li, P.S. (1988). *Ethnic inequality in a class society*. Toronto: Wall and Thompson.
McRoberts, H.A. (1985). Language and mobility: A comparison of three groups: The effects of origin. In M. Boyd, J. Goyder, E.F. Jones, H.A. McRoberts, P.C. Pineo, & J. Porter (Eds.), *Ascription and achievement: Studies in mobility and status attainment in Canada* (pp. 335-356). Ottawa: Carleton University Press.
McRoberts, H.A., Porter, J., Boyd, M., Goyder, J., Jones, F., & Pineo, P. (1976). Differencés dans la mobilité professionnelle des Francophones et des Anglophones. *Sociologie et Sociétés, 8*, 61-79.
Ornstein, M.D. (1981). The occupational mobility of men in Ontario. *Canadian Review of Sociology and Anthropology, 18*, 183-215.
Pedhazur, E.J. (1982). *Multiple regression in behavioral research* (2nd Ed.). New York: Holt, Rinehart & Winston.
Pineo, P.C. (1976). Social mobility in Canada: The current picture. *Sociological Focus, 9*, 109-123.
Pineo, P.C. (1985). Family size and status attainment. In M. Boyd, J. Goyder, F.E. Jones, H.A. McRoberts, P.C. Pineo, & J. Porter (Eds.), *Ascription and achievement: Studies in mobility and status attainment in Canada* (pp. 201-229). Ottawa: Carleton University Press.
Pineo, P., & Porter, J. (1985). Ethnic origin and occupational attainment. In M. Boyd, J. Goyder, F.E. Jones, H.A. McRoberts, P.C. Pineo, and J. Porter (Eds.), *Ascription and achievement: Studies in mobility and status attainment in Canada* (pp. 357-391). Ottawa: Carleton University Press.
Pineo, P.C., Porter, J., & McRoberts, H.A. (1977). The 1971 census and the socioeconomic classification of occupations. *Canadian Review of Sociology and Anthropology, 14*, 91-102.
Porter, J. (1965). *The vertical mosaic*. Toronto: University of Toronto Press.
Rashid, A. (1985). *The Muslim Canadians: A profile*. Ottawa: Minister of Supply and Services.
Reitz, J. (1980). *Survival of ethnic groups*. Toronto: McGraw-Hill Ryerson.
Richmond, A.H. (1967a). *Postwar immigrants in Canada*. Toronto: University of Toronto Press.
Richmond, A.H. (1967b). *Immigrants and ethnic groups in metropolitan Toronto*. Toronto: Institute for Behavioural Research, York University.
Richmond, A.H. (1973). Social mobility of immigrants. In J.E. Curtis & W.G. Scott (Eds.), *Social stratification: Canada* (pp. 174-186). Scarborough: Prentice-Hall of Canada.
Richmond, A.H., & Kalbach, W.E. (1980). *Factors in the adjustment of immigrants and their descendants*. (Cat. No. 99-761). Ottawa: Minister of Supply and Services.
Sandefur, G.D., & Scott, W.J. (1981). A dynamic analysis of migration: An assessment of the effects of age, family and career variables. *Demography, 18*, 355-368.
Satzewich, V., & Li, P.S. (1987). Immigrant labour in Canada: The cost and benefit of ethnic origin in the job market. *Canadian Journal of Sociology, 12*, 229-241.
Sharma, R.D. (1980). *Migration and fertility in a western Canadian metropolis*. Unpublished doctoral dissertation, University of Alberta.
Tepperman, L. (1975). *Social mobility in Canada*. Toronto: McGraw-Hill Ryerson.
Trovato, F., & Grindstaff, C.F. (1986). Economic status: A census of immigrant women at age thirty in Canada. *Canadian Review of Sociology and Anthropology, 23*, 569-587.
Wright, E.O. (1977). Marxist class categories and income inequality. *American Sociological Review, 42*, 32-55.
Wright, E.O. (1978). Race, class and income inequality. *American Journal of Sociology, 83*, 1368-1397.
Wright, E.O. (1979). *Class structure and income determination*. New York: Academic Press.

The Alberta Journal of Educational Research *Vol. XXXVIII, No. 3, September 1992, 219-234*

Estelle S. Gellman
Hofstra University

and

Mina Berkowitz
Long Island University

Factors Perceived as Important in Teacher Evaluation

The study sought to determine the factors that different segments of the school community consider important in evaluating a teacher. Elementary and secondary school teachers, novice and experienced teachers, student teachers, teacher educators, and parents were asked to rate the importance of 39 teacher characteristics. A factor analysis of their responses yielded three factors: (a) a responsiveness/attitude toward students factor, (b) a discipline/classroom management factor, and (c) a subject-matter expertise factor. Multivariate analyses of variance indicated that the responsiveness/attitude toward students factor was consistently rated as most important. The experienced teachers, the teacher educators, and the elementary school teachers rated the subject-matter expertise factor as significantly more important than the discipline/classroom management factor, while the reverse was true for the student teachers. No significant differences were found in the importance assigned to these two factors by the other groups. The findings indicate a common basis for teacher evaluation.

L'étude a tenté de déterminer les facteurs que différents segments de la communauté scolaire considèrent comme importants dans l'évaluation d'un maître. On a demandé à des maîtres des écoles élémentaires et secondaires, des maîtres novices et des maîtres expérimentés, des élèves-maîtres, des formateurs de maîtres, et des parents d'évaluer l'importance de 39 caractéristiques des maîtres. Une analyse factorielle de leurs réponses a révélé trois facteurs: (a) un facteur de réponse/attitude positive envers les élèves, (b) un facteur de discipline/gestion de classe et (c) un facteur d'expertise dans la matière d'enseignement. Des analyses multivariées de la variance ont indiqué que le facteur de réponse/attitude positive envers les élèves était toujours considéré comme le plus important. Les maîtres expérimentés, les formateurs de maîtres et les maîtres des écoles élémentaires ont évalué le facteur d'expertise dans la matière d'enseignement comme plus important, et cela d'une manière significative, que le facteur de discipline/gestion de classe, ce qui était le contraire pour les élèves-maîtres. Il n'y avait pas de différences significatives dans l'importance accordée à ces deux facteurs par les autres groupes. Les résultats indiquent une base commune pour l'évaluation des maîtres.

Estelle Gellman is a professor in the Department of Counseling, Research, Special Education and Rehabilitation where she teaches graduate courses in educational research, measurement, and statistics.
Mina Berkowitz is a professor of education at the Brooklyn Campus where she teaches courses in measurement and educational psychology in the undergraduate program in elementary education and the graduate programs in school psychology, special education, and elementary education.

Although there has been much interest in the importance of attracting good teachers into the classroom, there is not necessarily agreement as to what constitutes a "good" teacher (Epstein, 1985; Hamacheck, 1969; Issler, 1983; Tolor, 1973). Notwithstanding the large body of research investigating the relationship between various teacher behaviors and student achievement (Brophy & Good, 1986), the question remains as to whether the behaviors that result in improved student achievement are those that are perceived as characteristic of a good teacher. Furthermore, there may be different perceptions of good teaching among different segments of the school community. As indicated by Darling-Hammond, Wise, and Pease (1983), different evaluation constituencies—teachers, principals, parents, and public officials—have different stakes in the evaluation process. There is some indication that these different constituencies do in fact have different outlooks on the teaching profession and that these differences are reflected in differing perceptions of what constitutes a good teacher. It has been found, for example, that parent ratings of teacher effectiveness differ from those of faculty members and administrators (Epstein, 1985; Tolor, 1973). It may well be that even professional educators are not in agreement about what constitutes a good teacher. In a study investigating the importance that principals and secondary school biology teachers attach to different criteria of teacher effectiveness, Searles and Ng (1982) found that characteristics dealing with subject-matter preparation and enthusiasm for the subject were among the most highly rated criteria by both groups. In Epstein's (1985) study of elementary school teachers, on the other hand, it was found that neither principals' nor parents' ratings of teachers were correlated with advanced study in the subjects taught. These differences suggest that there may be systematic variation in the way that different segments of the educational community conceptualize a good teacher. If there are different perceptions of good teaching, questions may arise as to the criteria to be used in assessing teacher performance.

In reviewing the literature on teacher evaluation, Strike and Millman (1983) suggest that, in order for assessment to be successful, teachers and evaluators must have common views concerning education, evaluation, and the role of the teacher. Kauchak, Peterson, and Driscoll (1985) further suggest that professionalism demands that teacher perspectives be considered in any evaluation system. Selmes (1986) goes so far as to suggest that teacher evaluation should be based on a new set of assumptions, including the assumption that evaluation criteria should be jointly negotiated and agreed upon by teachers and evaluators.

If successful evaluation procedures require agreement on the criteria of evaluation, then it is important to know the extent to which different segments of the school community agree on the characteristics to be assessed. Different perceptions of a good teacher may be held by elementary and secondary school teachers, teachers differing in their degree of experience, student teachers, and teacher educators. Parents and teachers may also differ in their perceptions of good teaching. These different groups may differ not only in those characteristics that are identified as being important, but also in the degree of importance ascribed to particular characteristics.

In order to explore these issues, the aim of the present study was to first identify the factors perceived as important in evaluating a teacher. Although the literature on teacher effectiveness identifies a variety of discrete behaviors related to student outcomes, the purpose of the present study was to determine whether in examining perceptions of good teaching these characteristics could be explained by fewer more general factors. That such is often the case is indicated by factor analytic studies of teacher evaluation instruments. Swartz, White, Stuck, and Patterson (1990), for example, in examining the factorial structure of 27 teaching practices on the North Carolina Teaching Appraisal Instrument (TPAI), found that the items on the scale were most parsimoniously described by two factors, one relating to the teacher's presentation of the lesson and the other relating to the teacher's management of the class. Similarly, Teddlie, Virgilio, and Oescher (1990) found that the 35 items on the Virgilio Teacher Behavior Inventory (VTBI) could be explained by five factors: one classroom management, two instruction, and two classroom climate dimensions.

The second part of the present study then examined the importance ascribed to each of the emerging factors. The overall importance rating of each factor was examined as well as the relative importance ascribed to each of the factors by different segments of the school community: parents, teachers (broken into subgroups of novice, experienced, elementary, and secondary teachers), student teachers, and teacher educators.

Method

In order to identify those factors that different groups perceive as important in evaluating a teacher, a questionnaire was developed that asked respondents to indicate the importance that they would ascribe to each of 39 teacher characteristics.

Construction of the Instrument

A preliminary list of teacher characteristics was derived from a review of the literature on teacher evaluation. Additional characteristics were then obtained by asking teachers, parents attending PTA meetings, and high school students attending classes at two school districts in Nassau County to list the five characteristics that they would consider most important in evaluating a teacher. The items were originally grouped into four categories: (a) classroom management/climate, (b) personal qualities, (c) subject matter competence/communication skills and (d) interpersonal relationships. Feedback from a small group of educators indicated, however, that these categories were insufficient; several of the items did not appear to be appropriately placed in any of these categories. The experimenters then reclassified the items and assigned category names to similarly grouped items. It was felt that the items were best described by the following seven category names: (a) discipline, (b) classroom management, (c) subject matter competence, (d) attitude toward students, (e) demands on students, (f) pedagogical skills, and (g) personal characteristics.

A preliminary scale was then developed which included between seven and 15 items from each category. Where a category had fewer than seven items, additional items were written by the investigators. These items were

then distributed to 158 teachers and graduate education students with instructions to indicate the category represented by each item.

A seven-point Likert-type scale of 39 items was then developed that included those items (a) on which there was over 48% agreement on the category assignment, and (b) that were not assigned to any other category more than 32% of the time. The mean percent agreement for the items in each category ranged from 54.43% to 92.7%, with a mean for the entire scale of 75%. Three forms of the final scale were constructed, each form containing a different random order of the items. Respondents were instructed to indicate the degree of importance that they would assign to each of the 39 characteristics in evaluating a teacher by circling the appropriate scale value. Values ranged from 1 through 7, with a value of 1 indicating that the characteristic was considered a major factor in evaluating a teacher and a value of 7 indicating that the characteristic was considered to be totally irrelevant.

Sample and Procedure

The scale was administered to 488 parents and educators in the New York metropolitan area, with random assignment of the different forms. Subjects were sought at PTA meetings, professional meetings, graduate education classes, and university departments of education. Respondents were assured of anonymity and were requested to complete a personal data form as well as to rate the importance of each of the items on the scale of teacher characteristics.

As noted in Table 1, the sample consisted of 102 males and 374 females, with 12 respondents not completing this item. Included in the sample were 30 teacher educators, 53 student teachers, and 64 parents who were not teachers. The sample also included 120 elementary school teachers, defined as those

Table 1
Number of Respondents in each Category Broken Down by Sex and Experience

		Years of Experience														
		1-5			*6-10*			*11 or more*			*Missing*			*Total*		
Category	n	M	F	NR	M	F	NR	M	F	NR	M	F	NR	M	F	NR
Elementary Teachers	120	4	45	0	2	20	0	5	19	0	3	20	2	14	104	2
Secondary Teachers	100	11	20	0	9	10	0	20	14	2	3	11	0	43	55	2
Teacher Educators	30	0	1	0	1	1	0	9	10	0	2	5	1	12	17	1
Student Teachers	53	—	—	—	—	—	—	—	—	—	—	—	—	5	45	3
Parents	64	—	—	—	—	—	—	—	—	—	—	—	—	9	54	1
Total Sample	488	—	—	—	—	—	—	—	—	—	—	—	—	102	374	12

Note. NR represents individuals who did not indicate their sex. The subgroup totals do not equal the total sample because not all individuals could be assigned to a subgroup. Dashes indicate the absence of respondents in the category.

who taught in grades N-6 and had no junior high school, high school or college teaching experience, and 100 secondary school teachers, defined as those who taught in grades 7-12 and had no experience at the elementary or nursery school level. Teachers who had experience at both the elementary and secondary levels were not included in either of these subgroups. In the sample of teachers, 121 were new teachers, defined as having had five or fewer years of experience, and 108 were experienced teachers, defined as having had 11 or more years of experience.

Data Analysis

Part 1. Part 1 of the study was designed to identify the factors that are perceived as important in evaluating a teacher.

An initial iterative principal axis factor analysis was performed on the importance ratings of those respondents who completed all of the items on the scale. The purpose of this analysis was to ascertain how many factors would best describe the data. The criteria entering into this decision included (a) the initial eigenvalues (i.e., not subjected to iteration), (b) a scree examination, and (c) the percent of variance explained by the factor. Based on these criteria, the three- and four-factor solutions were further examined for interpretability and were subjected to a principal axis factor analysis with a Varimax rotation. This analysis guided the final judgment as to the number of factors that best explained the data and was used for the development of factor-based scales (Kim & Mueller, 1978, p. 70). SPSS, Release 4.0 for the IBM VM/CMS, was used for all data analyses.

Part 2. Part 2 of the study examined the relative importance assigned to the factors identified in Part 1.

In order to compare the importance ascribed to the factors by the different groups, a factor-based scale was developed for each factor. For an item to be included on a scale, it had to load .4 or higher on the rotated factor in the three-factor solution and have a loading less than .35 on the other two factors.

Based on the scores on the factor-based scales, a repeated measures multivariate analysis of variance was performed to compare the importance ascribed to each of the factors by the total sample. Then four profile analyses were performed to compare the ratings assigned by different subgroups to the different factors. The groups compared were (a) elementary and secondary school teachers, (b) new and experienced teachers, (c) parents and experienced teachers, and (d) student teachers, teacher educators, and experienced teachers.

Results

Part 1

As reported in Table 2, the initial factor analysis yielded four factors with eigenvalues greater than 1, with the fourth factor barely meeting this criterion. A scree examination yields an "elbow" that falls to the left of the fourth factor, with the eigenvalues clearly leveling off by the fourth factor. As may be also noted in Table 2, the first three factors explained 79.21% of the variance, with the fourth factor explaining another 6.92%.

Based on the above, both the three- and the four-factor solutions were further examined for interpretability. Inspecting the rotated factor matrices, as

Table 2
Initial Eigenvalues and Percent of Variance Explained by Each Factor in the Iterative Factor Analysis

Factor	*Eigenvalue*	*Percent of Variance Explained*
1	7.762	49.55
2	2.730	17.43
3	1.916	12.23
4	1.083	6.91
5	.976	6.23
6	.702	4.48
7	.674	4.30
8	.491	3.13
9	.441	2.82
10	.356	2.27
11	.283	1.81
12	.254	1.62
13	.242	1.54
14	.178	1.14
15	.143	.91
16	.121	.77
17	.100	.64
18	.069	.44
19	.053	.34
20	.022	.14
21	.006	.04
22	−.003	<.01
23	−.023	<.01
24	−.054	<.01
25	−.081	<.01
26	−.100	<.01
27	−.102	<.01
28	−.133	<.01
29	−.141	<.01
30	−.154	<.01
31	−.179	<.01
32	−.196	<.01
33	−.204	<.01
34	−.230	<.01
35	−.240	<.01
36	−.246	<.01
37	−.263	<.01
38	−.275	<.01
39	−.314	<.01

reported in Table 3, it may be seen that only five items had factor loadings of .4 or higher on the fourth factor and that none of them had loadings greater than .61. Furthermore, one of these items (item 6) had a higher loading on Factor 2, and two of the items had loadings in the area of .30 on either Factor 1 or Factor 2.

Looking further at the item correlation matrix, it was found that the intercorrelations among the items with high loadings on the fourth factor ranged from .17 to .47, with a mean of .31. Many of the correlations between the items

Table 3
Rotated Factor Matrices for the Three and Four-Factor Solutions

	Factor Loading						
	Three Factor Solution			Four Factor Solution			
Item	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*
1	.37787	.12822	.05616	.29351	.03029	.04155	.43555
2	.03216	.20330	.37192	.03467	.20968	.37103	.02006
3	.54809	−.06066	.01172	.54761	.05486	.00996	.06112
4	.25341	.11860	.54135	.22137	.08851	.53754	.18337
5	.20224	.60761	.04420	.15928	.56919	.03586	.24887
6	.26692	.51790	.15474	.17896	.43854	.14062	.43644
7	.21181	.06772	.60636	.23047	.09260	.61053	−.02949
8	.60375	−.04192	.03989	.59663	−.04228	.03721	.09546
9	.06219	.22293	.15342	.05184	.21475	.15146	.07127
10	.00763	.11663	.51102	.00199	.11237	.50812	.04980
11	.07561	.62401	.11678	.07801	.63982	.11304	.04980
12	.54125	.04628	.01246	.52594	.03745	.00824	.12789
13	.56643	.07550	.13017	.57631	.09283	.12845	.04171
14	.06421	.49383	.12093	.03084	.46278	.11532	.18617
15	−.02195	.14716	.01376	−.00627	.16575	.01589	−.05474
16	.51721	.20751	.14141	.48230	.17847	.13486	.22097
17	−.01390	.53533	.04573	−.03197	.51939	.04213	.11824
18	.55789	.03356	.05478	.54619	.02775	.05098	.11730
19	.14747	.69800	.05105	.13257	.69209	.04484	.13619
20	.37410	.18125	.13908	.27016	.06111	.12605	.53259
21	.61560	.17338	.06788	.57830	.14249	.06045	.23772
22	.30035	.46504	.19576	.30749	.48523	.19335	.04993
23	.42487	.21420	.21462	.46066	.25996	.21766	−.05908
24	.41924	.09317	.09300	.39677	.07501	.08882	.14896
25	.40021	.20341	.12043	.42256	.23480	.12019	−.01676
26	.34906	.31430	.20562	.22670	.18816	.19533	.60529
27	.62666	.11790	.08445	.62495	.12526	.08105	.09383
28	.49068	.40640	.08497	.48884	.41626	.08029	.10289
29	.68557	.09135	−.04066	.65535	.06819	−.04672	.20642
30	.05984	.55726	.19792	.06296	.57167	.19571	.04214
31	.27614	.39637	.11417	.18170	.30715	.09944	.45845
32	.55723	.14273	−.08793	.52726	.11911	−.09358	.19315
33	.52809	−.03208	.10834	.57089	.01272	.11151	−.08880
34	.18041	.07226	.68187	.18065	.07764	.67882	.04538
35	.11872	.62133	.01942	.12142	.63800	.01471	.05194
36	.50744	.13048	.16616	.49601	.12684	.16278	.11817
37	.02278	.08620	.74299	.02070	.08724	.73993	.03945
38	−.00695	.10324	.45496	−.05296	.05503	.45579	.20096
39	.17137	.20968	.18934	.20967	.25930	.19436	−.11265

with high loadings on Factor 4 and the items with high loadings on Factor 1 or Factor 2, however, fell in the same range. Additionally, as may be noted in Table 3, every item on Factor 4 had a loading greater than .30 on either Factor 1 or Factor 2 in the three-factor solution, with item 6 having a higher loading on Factor 2 in the three-factor solution than it did on Factor 4.

Considering the substance of the items with high loadings on Factor 4, they might be broadly interpreted as representative of a personal qualities factor,

including such items as (a) Has a good personality, (b) Has a pleasant voice, and (c) Has a sense of humor. These items, however, do not seem to meaningfully fit with item 31, Displays good use of bulletin board space, which has the third highest loading on the factor. In the three-factor solution, the factor loading of item 31 increases on Factor 2, which appears to be related to classroom management, and the other items increase their loadings on Factor 1, which appears to reflect the teacher-student relationship. Taking these considerations into account, as well as the question of the reliability of a factor-based scale including those few high-loading items on Factor 4, only three factors were used for subsequent analyses. The items included on the three factor scales are presented in Table 4. Nine of the items on the original scale were not included on any of the factor-based scales.

Those items with high factor loadings on each of the factors were examined in order to define the three factors. Factor 1 was defined as a responsiveness/attitude toward students factor, Factor 2 as a discipline/classroom management factor, and Factor 3 as a subject-matter expertise factor.

Table 4
Items Included on each Factor Scale

Factor 1	*Factor 2*	*Factor 3*
Respects the students (#3)	Attendance is taken regularly (#5)	Reads extensively in the subject area (#4)
Listens to students (#8)	Is neatly dressed (#6)	Knows subject in depth (#7)
Is patient (#12)	Is a strict disciplinarian (#11)	Is an acknowledged expert in the subject taught (#10)
Admits mistakes honestly (#13)	Does not allow talking during instruction (#14)	Has taken advanced courses in subject area (#34)
Asks questions in class (#16)	Students are punished for being late (#17)	
Is enthusuastic (#18)	Maintains a strict set of rules to be obeyed (#19)	Has an advanced degree in the subject area (#37)
Uses a variety of teaching skills (#21)	Keeps accurate records (#22)	Has professional publications in subject area (#38)
Gives fair exams (#23)	Regularly gives quizzes and exams (#30)	
Respects students' privacy (#24)	Misbehavior is swiftly punished (#35)	
Sets demands on level of students (#25)		
Provides feedback (#27)		
Makes students feel important (#29)		
Knows each student's level of competence (#32)		
Stimulates intellectual curiosity (#33)		
Repeats material as needed (#36)		

Part 2.

Total group analysis. When looking at the mean importance ratings assigned to the three factors by the total group, as reported in Table 5, it was found that the responsiveness/attitude toward students factor was rated as more important than the subject-matter expertise factor, and that the subject-matter expertise factor was rated as more important than the discipline/classroom management factor. The overall multivariate analysis of variance was significant; Wilks' λ=.189, $F(2,482)$=1,033.55, p<.05. It may also be noted, as reported in Table 6, that the difference in the importance ascribed to the responsiveness/attitude toward students factor and the subject-matter expertise factor was also significant, as was the difference between the importance ratings of the subject-matter expertise factor and the discipline/classroom management factor.

Subgroup analyses. As also noted in Table 5, every subgroup rated the responsiveness/attitude toward students factor as most important and, as may be further noted in Table 6, the responsiveness/attitude toward students factor was rated by every group as significantly more important than the next highest rated factor.

The subject-matter expertise factor was rated as more important than the discipline/classroom management factor by all subgroups except the parents and the student teachers. As may be noted in Table 6, the difference in the importance ratings of these two factors was significant for all subgroups except for the: (a) new teachers, (b) secondary school teachers, and (c) parents.

Examining the pattern of sub-group response across the factors, the profile of the elementary school teachers was significantly different from that of the

Table 5
Mean Importance Ratings Assigned to the Three Factors by the Different Groups

		Factor					
		Responsiveness to Students		*Discipline/ Management*		*Subject-Matter Expertise*	
Group	*n*	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*	*Mean*	*Standard Deviation*
Parents	64	1.75	.53	3.03	.91	3.22	.94
Experienced Teachers (All)	108	1.85	.62	3.69	1.14	3.15	.97
Experienced Teachers (Excluding teacher educators)	89	1.94	.63	3.77	1.17	3.19	.98
New Teachers	121	1.71	.45	3.38	.85	3.22	.95
Elementary Teachers	118	1.71	.46	3.61	1.00	3.31	1.00
Secondary Teachers	100	1.79	.51	3.30	1.05	3.27	.92
Teacher Educators	29	1.51	.36	3.52	.98	3.07	.91
Student Teachers	52	1.71	.53	3.23	.82	3.70	1.02
Total Sample	484	1.75	.54	3.42	1.00	3.26	.97

Note. Where n is less than the number of such individuals in the sample, it is because respondents who did not respond to every item on the scale were not included in this analysis.

Table 6
Significance of the Difference in the Importance Ratings Assigned to the Three Factors by the Groups Compared in the Profile Analyses

			Comparison					
			Factor 1 vs Factor 2		*Factor 1 vs Factor 3*		*Factor 2 vs Factor 3*	
Group	n	df	F	*p*	F	*p*	F	*p*
Total Sample	484	1,483	1,478.53	.000	1,128.71	.000	8.37	.004
Teacher Educators	29	1,167	121.27	.000	77.91	.000	4.95	.027
Student Teachers	52	1,167	126.07	.000	227.54	.000	9.75	.002
Experienced Teachers (Excluding Teacher Educators)	89	1,167	309.36	.000	152.79	.000	25.93	.000
Experienced Teachers (Including Teacher Educators)	108	1,227	392.29	.000	187.02	.000	22.24	.000
New Teachers	121	1,227	365.26	.000	287.22	.000	2.15	.144
Elementary Teachers	118	1,216	440.71	.000	330.02	.000	7.28	.008
Secondary Teachers	100	1,216	236.31	.000	239.55	.000	.08	.779
Parents	64	1,170	101.68	.000	155.86	.000	2.07	.152
Experienced Teachers (Including Teacher Educators)	108	1,170	356.55	.000	205.08	.000	28.21	.000

Note. Where n is less than the number of such individuals in the sample, it is because respondents who did not respond to every item on the scale were not included in this analysis.

secondary school teachers (λ=.962, $F(2,215) = 4.27$, $p<.05$). Although both groups rated subject-matter expertise as more important than discipline/classroom management, the difference in the ratings of these two factors was significant for the elementary school teachers but not the secondary school teachers. It may also be noted that the profile of the parents was significantly different from that of the experienced teachers (λ=.892, $F(2,169)=10.21$, $p<.05$). While the experienced teachers rated subject-matter expertise as significantly more important than discipline/classroom management, this pattern was not duplicated for the parents; although the parents reversed the relative importance ascribed to discipline/classroom management and subject-matter expertise, the difference between these two factors was not significant for the parents.

A significant difference was also found in the profile analysis comparing the student teachers, teacher educators and experienced teachers (λ=.820, $F(4,334)= 8.67$, $p<.05$). The analysis did not indicate any significant difference among the groups in the degree to which Factor 1 was seen as more important than Factor 2 ($F(2,167)=2.59$, $p.>05$). Differences were found, however, in the differentiation between Factors 2 and 3 ($F(1,167)=16.18$, $p<.05$). Contrasts indicated that although the experienced teachers and the teacher educators did not

significantly differ in the extent to which they differentiated among Factors 2 and 3 ($F(1,167)=.35$, $p>.05$), there was a significant difference between these two groups and the student teachers ($F(1,167)=26.98$, $p<.05$). As may be noted in Table 5, this difference reflects the fact that, while the student teachers rated the discipline/classroom management factor as significantly more important than the subject-matter expertise factor, the reverse was true for the experienced teachers and teacher educators.

No significant profile differences were found when comparing the importance ascribed to the three factors by the new and experienced teachers.

Discussion

Although the present study was based on responses to a questionnaire designed to include items relating to seven different aspects of teaching cited in the literature, the factor analysis did not yield seven distinct factors. Only three clear factors were found to underlie the perception of what constitutes a good teacher: (a) a responsiveness/attitude toward students factor, (b) a discipline/classroom management factor, and (c) a subject-matter expertise factor.

The responsiveness/attitude toward students factor emerged first and was most diffuse. This factor seems to reflect aspects of both personal and pedagogical responsiveness, as well as general personal and attitudinal characteristics. Most of the high-loading items on this factor had been assigned to four of the seven categories on which the questionnaire was based: (a) attitude toward students, (b) demands on students, (c) pedagogical skills, and (d) personal characteristics. It is quite clear, however, that these items did not reflect four distinct factors.

The discipline/classroom management factor was the second factor to emerge. In constructing the survey instrument, we assigned most of the items with high loadings on this factor to what appeared to be two distinct categories, discipline and classroom management. The results of the factor analysis indicated, however, that items from both categories were highly intercorrelated and had high loadings on a single factor. This factor was much more sharply defined than the first factor, with only one high-loading item, "Is neatly dressed," not clearly reflecting a classroom management and/or discipline function. One might, however, interpret this item to be reflective of self-discipline on the part of the teacher.

The third factor was also very sharply defined, with every high-loading item clearly reflecting advanced training or expertise in the subject-matter being taught. Every item with a high loading on this factor had been originally assigned to the category of subject-matter expertise, with only one questionnaire item originally assigned to this category not meeting the requirements for inclusion on the factor-based scale for the third factor. This item, "Is able to answer most questions without a book," had a factor loading of .37 on the factor.

The factors emerging in the present study are not surprising in view of the Swartz et al. (1990) and Teddlie et al. (1990) studies of teacher appraisal instruments. Although the factor structure is not identical to that found in the other studies, there is considerable similarity even though these other studies investigated the structure of the practices assessed on actual appraisal instru-

ments whereas the present study focused on the importance ascribed to various teacher behaviors.

Classroom management, for example, emerged as a single factor in the present study as well as in both the Teddlie et al. (1990) and the Swartz et al. (1990) studies. Of particular interest is the fact that, although the TPAI (Swartz et al., 1990) yields separate scores for management of instructional time and management of student behavior, the investigators concluded on the basis of a factor analysis of teacher ratings that a two-factor solution in which aspects of both types of behavior loaded on a single factor was most parsimonious in describing the data. The instrument constructed for the present study was also designed with separate categories for discipline and classroom management, and the results of the factor analysis similarly indicated that these practices formed a single factor.

The second factor emerging from the analysis of the TPAI was defined as an instructional presentation factor on which the high loading practices parallel many of the high-loading items found on the responsiveness/attitude toward students factor in the present study. The other factors emerging from the analysis of the VTBI (Teddlie et al., 1990) included two instruction factors and two classroom climate factors. Although the present study did not find distinct factors that were definable in this way, many of the items with high loadings on Factor 1 in the present study do reflect classroom climate or instruction. The question is whether these practices can be described by a single underlying factor as found in the present study and the Swartz et al. (1990) study or whether more factors are needed to meaningfully describe the data.

It is also of interest to note that neither the Swartz et al. (1990) nor the Teddlie et al. (1990) studies yielded anything reflecting the subject-matter expertise factor emerging in the present study. This finding is understandable because both the TPAI and the VTBI are essentially observation instruments that focus on observable teacher behaviors. The factors emerging from factor analytic studies of any assessment instrument are necessarily restricted by the nature of the items being assessed, and a subject-matter expertise factor would not emerge in a study where this factor is clearly not assessed. The present study included items relating to this variable, and these items clearly formed a factor distinct from the discipline/classroom management and responsiveness attitude toward students factor.

Interestingly, these latter two factors are also congruent with the research on teacher effectiveness. The research of Soar and Soar (cited in Brophy & Good, 1986), for example, indicates that emotional climate and teacher management are independent factors affecting achievement. Their definition of emotional climate includes many of the characteristics included on the first factor emerging in the present study, whereas their definition of teacher management subsumes both the classroom management and discipline aspects of Factor 2. In the Brophy and Good review (1986), in fact, classroom management was repeatedly cited as a factor related to student achievement and, although the findings are more mixed, several of the studies reviewed also found student achievement to be related to teacher behaviors that had high

loadings on the responsiveness/attitude toward students factor emerging in the present study.

The difference between the factors that were cited in the Brophy and Good review (1986) as related to student achievement and those that the present study found to be considered important in teacher evaluation was that (a) the subject-matter competence of the teacher was not cited as a factor in the studies on student achievement, and (b) many of the specific pedagogical techniques found to be related to student achievement were not included on any of the factors in the present study. Although the questionnaire used in the present study did include general items relating to such areas as feedback, questioning, demands on students, and setting levels of expectation, it did not include items relating to specific classroom techniques such as grouping, rephrasing, wait time, time on task, pacing, providing an overview or summary, et al.

Those items that did relate to more general pedagogical techniques tended to load on the responsiveness/attitude toward students factor. It would appear that such teacher behaviors as questioning, providing feedback, and adjusting demands to the level of the students are perceived within the general factor of responsiveness to student learning needs. Because the present study did not include the more specific behaviors, it is not clear whether they also would be seen in this way. Further investigation is needed to clarify this question, but the results of the present study suggest that many of the pedagogical practices related to student achievement may have a common thread of reflecting a responsiveness to student needs. It seems that this factor reflects more than positive affect on the part of the teacher and also includes a recognition of the cognitive needs of the student.

The third factor emerging in the present study, subject-matter expertise, has been only tangentially mentioned in the research on teacher effectiveness. One explanation may be that the research in this area parallels assessment instruments such as the TPAI and the VTBI in focusing on observable teacher behaviors. Issues such as previous schooling or involvement with out-of-class professional activities are not directly observable, and it may be assumed that these factors are a prerequisite to such variables as clarity of presentation or fluency. The present study does indicate, however, that subject-matter expertise, whether assumed to be prerequisite to other teacher behaviors or not, is a separate and distinct factor to be considered in evaluating a teacher.

In examining the importance ascribed to the three factors emerging in the present study, it is of particular interest to note the high degree of consistency among the different segments of the school community in rating the responsiveness/attitude toward students factor as most important. Not only was the importance ascribed to the discipline/classroom management and subject-matter expertise factors consistently lower than that ascribed to the responsiveness/attitude toward students factor, but in every instance these two factors were also seen as closer in importance to each other than either was to the responsiveness/attitude toward students factor.

Although the analysis of the responses of the total group indicated that significantly more importance was ascribed to the subject-matter expertise factor than to the discipline/classroom management factor, this pattern was

not consistent among the subgroups. Whereas the teacher educators, experienced teachers, and elementary school teachers rated the behaviors on the subject-matter expertise factor as more important than those on the discipline/classroom management factor, the reverse was true for the student teachers. The parents, the new teachers, and the secondary school teachers did not rate these two factors as differing significantly in importance.

The consistency of this pattern is borne out in the profile analyses that were performed to compare the pattern of response among student teachers, teacher educators, and experienced teachers and between new and experienced teachers, elementary and secondary teachers, and parents and experienced teachers. In those instances where profile differences were found, the difference could be attributed to the fact that the groups compared differed in the relative importance ascribed to Factors 2 and 3.

Although the general pattern was to rate the subject-matter expertise factor as more important than the discipline/classroom management factor, it is interesting to note that the two groups in which this pattern was not found were the parents and the student teachers, the two groups without any or with minimal teaching experience. Also, there was no significant difference in the importance that the new teachers ascribed to these two factors. One may speculate that experienced teachers who know that they are capable of handling a class of students may be less concerned with issues of discipline than with subject-matter competence where, in order to remain current in the field, they must continuously update their knowledge and skill. Student teachers, on the other hand, are in the process of learning the latest developments in the field, but have not had the experience of testing their own competence in disciplining and managing a class. They may therefore perceive this task as more challenging and see success in this areas as more important.

It is also interesting to note that, while the elementary school teachers rated the subject-matter expertise factor as more important than the discipline/classroom management factor, there was no difference in the importance ascribed to these two factors by the secondary teachers. Given the different teaching environments of elementary and secondary school teachers, one might expect secondary school teachers to focus more on subject-matter expertise. The fact that this factor was not rated as more important may suggest that elementary school teachers assume that most of their colleagues have the minimum management abilities required and have more concern about teachers remaining current in the subject areas being taught than in their classroom management abilities. Secondary school teachers, on the other hand, are dealing with an age group that can be extremely disruptive in the absence of appropriate management and discipline practices and may consequently put as much emphasis on this factor as on subject-matter expertise.

These findings clearly indicate that the way teachers respond to students is perceived as being of primary importance among all segments of the school community, and that the attitude the teacher displays toward students is perceived as more important than either the depth of knowledge in the subject being taught or classroom management skills. It should be noted, however, that all the factors were considered important. The mean on each of the factor scales fell on the positive side of the scale, where a rating of 1 indicated that the

behavior was considered a major factor in teacher evaluation. While the responsiveness/attitude toward students factor consistently fell between the most positive values of 1 and 2 on a 7-point scale, however, the mean of the other two factor scales fell between 3 and 4.

The importance that is ascribed to this factor clearly indicates that teacher evaluation must focus on ways to assess those complex and relatively intangible behaviors that reflect the quality of the interaction between teacher and student and appear to be perceived as the foundation of good teaching. Although this interaction may be difficult to assess and may not be necessarily related to the conventional indices of student achievement, it must be included in any teacher assessment process designed to reflect those aspects of teaching that teachers themselves consider important.

If, as suggested by Strike and Millman (1983), there must be agreement between teacher and evaluator for assessment to be successful, the strong agreement that was found among the different segments of the school community indicates that successful evaluation is possible. According to Darling-Hammond, Wise and Pease (1983), because approaches to teacher evaluation vary according to those teacher attributes believed to be important for effective teaching, it is clear that any approach to evaluation must address the issue of how to assess those less tangible behaviors that define the nature of the teacher's attitude and relationship to students.

Acknowledgments

We gratefully acknowledge the assistance of Liora Schmelkin in analyzing and interpreting the data and of Marci Mann in searching the literature, in recording and entering the data on the computer, and in executing the computer analyses.

Portions of this research were presented at meetings of the Northeastern Educational Research Association and the Eastern Educational Research Association.

References

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 328-375). New York: Macmillan.

Darling-Hammond, L., Wise, A.E., & Pease, S.R. (1983). Teacher evaluation in the organizational context. *Review of Educational Research, 53*(3), 285-329.

Epstein, J.L. (1985). A question of merit: Principals' and parents' evaluations of teachers. *Educational Researcher, 14(7), 3-10.*

Hamachek, D. (1969). Characteristics of good teachers and implications for teacher education. *Phi Delta Kappan, 50,* 341-344.

Issler, K. (1983). A conception of excellence in teaching. Education, 103, 338-343.

Kauchak, D., Peterson, K., & Driscoll, A. (1985). An interview study of teachers' attitudes toward teacher evaluation practices. *Journal of Research and Development in Education, 19,* 32-37.

Kim, J., & Mueller, C.W. (1978). *Factor analysis: statistical methods and practical issues* (Sage University paper series on Quantitative Applications in the Social Sciences, 07-014). Beverly Hills: Sage.

Searles, W.E., & Ng, R.W.M. (1982). A comparison of teacher and principal perception of an outstanding biology teacher, *Journal of Research in Science Teaching, 19,* 487-495.

Selmes, A. (1986). Teacher evaluation in the classroom. *Educational Management and Administration, 14,* 191-196.

Strike, K.A., & Millman, J. (1983). Nontechnical questions about teacher evaluation systems in elementary and secondary schools: A research agenda. *Educational Evaluation and Policy Analysis, 5,* 389-397.

Swartz, C.W., White, K.P., Stuck, G.B., & Patterson, T. (1990). The factorial structure of the North Carolina Teaching Performance Appraisal Instrument. *Educational and Psychological Measurement, 50,* 175-182.

Teddlie, C., Virgilio, I., & Oescher, J. (1990). Development and validation of the Virgilio Teacher Behavior Instrument, *Educational and Psychological Measurement, 50,* 421-430.

Tolor, A. (1973). Evaluation of perceived teacher effectiveness. *Journal of Educational Psychology, 64,* 98-104.

The Alberta Journal of Educational Research *Vol. XXXVIII, No. 3, September 1992, 235-250*

Dushyanta Persaud
Winnipeg School Division No. 1

and

Paul R. Madak
University of Manitoba

Graduates and Dropouts: Comparing Perceptions of Self, Family, and School Supports

A survey approach was used to compare 92 students about to graduate from three Winnipeg high schools with 38 dropouts. Graduates were obtained from an alternative school, a vocational school and a school geared towards academic programs. Dropouts were recruited from these schools and organizations providing upgrading programs. Results revealed that graduates perceived themselves as being smarter, more ambitious and responsible, more involved in church and volunteer work, less involved with drugs/alcohol, and less involved in criminal activities than dropouts. Furthermore, graduates were more satisfied with their school experiences and felt parents and teachers were more supportive.

Une approche employant le questionnaire était utilisée pour comparer 92 étudiants finissants du secondaire à 38 étudiants qui ont abandonné leurs études. Les étudiants finissants provenaient d'une école spéciale, d'une école professionnelle et d'une école fondée sur les programmes académiques. Les étudiants qui ont abandonné provenaient d'écoles et d'organisations délivrant des programmes de rattrapage. Les résultats ont révélé que les étudiants finissants se considèrent plus intelligents, plus ambitieux et responsables, plus impliqués à l'église et au travail volontaire, moins reliés aux drogues et à l'alcool et moins impliqués dans des activités criminelles que les étudiants qui ont abandonné leurs études. En outre, les étudiants finissants étaient plus satisfaits avec leurs expériences à l'école et trouvaient plus positifs les parents et les professeurs.

One of the hottest topics in education today is the issue of early school leavers, or as they are commonly referred to, high school dropouts. Over the last seven to 10 years much research has been carried out in the areas of high school dropouts and students who are at risk of dropping out of school (Mann, 1986a; Morris, Pawlovich, & McCall, 1991; Rumberger, 1987; Strother, 1986; Wehlage, Rutter, Smith, Lesko, & Fernandez, 1989). Although the dropout rate of approximately 30% in Canada has remained relatively constant over the last 10 years (Morris et al., 1991), concern about the effects of early school withdrawal has resulted in the highly publicized Canada Employment and Immigration's Stay-In-School Program initiative. This concern has also led to a search for the factors that render students at risk of dropping out prematurely so as to

---

Dushyanta Persaud teaches mathematics and science at Sisler high school.
Paul Madak is an associate professor in the Department of Educational Psychology. His research interests include school dropouts, sexual abuse prevention, and program evaluation.

identify them and implement prevention/intervention strategies. The purpose of this study was to carry out a direct comparison of individuals who had dropped out of three Winnipeg high schools with individuals who had remained in school until they had graduated, in order to provide educators with additional information that might assist in the reduction/prevention of dropout behavior.

Students who leave school before obtaining a high school diploma place themselves at risk to a lifetime of personal and career problems, as well as creating a problem for Canadian society at large. Dropping out not only reduces the probability that the individual will be able to find full-time employment, but also increases the probability that they will be an added burden to the Canadian tax payer. It has been estimated that unemployed dropouts cost the government billions of dollars in terms of lost tax revenue. Furthermore, there are the added costs associated with the fact that school dropouts are more likely than graduates to need unemployment insurance and/or welfare, be involved with the justice system, and require more health care and other social services (Morris et al., 1991; Rumberger, 1987; Wehlage et al., 1989). The Ontario Ministry of Education reported that in 1981 20% of the dropouts in the 15-24 age group were unemployed, whereas only 1% of the graduates were unemployed (Karp, 1988). Furthermore, a second study in Ontario indicated that the average income of dropouts was approximately $3,770.00 less per year than for graduates (Sullivan, 1988).

The number of individuals who drop out of Canadian schools before obtaining a high school diploma has been reported to vary greatly from province to province and from region to region within provinces. In fact, the dropout rate can vary from a low of 5% to a high of 80% (Morris et al., 1991). This large variation is partly due to the lack of consistent data collection processes across and within the provinces, and partly because the many regional differences cause the rates to vary.

The typical high school dropout has been described in the literature as being predominantly male, from a low socioeconomic and single parent family, having a poor attendance and academic record, above average in age for his or her grade level, not involved in school activities, belonging to a language minority group, and having changed schools a number of times (Beck & Muia, 1980; Hammons, 1988; Mincer, 1989; Morris et al., 1991; Strother, 1986; Wehlage et al., 1989). However, from a review of the research literature on school dropouts, it is clear that no single factor, or group of factors, has been identified as the cause behind an individual's decision to leave school before obtaining a high school diploma. Rather, dropping out of school appears to be a process that occurs over time and is made up of a series of events in the individual's school experience (Finn, 1989; Morris et al., 1991; Rumberger, 1987; Wehlage et al., 1989). Leaving school before obtaining a high school diploma appears to result from a complex interaction between variables and events that begin as soon as the student enters school.

Why Students Drop Out

The reasons for dropping out of school can be broken down into three main categories: (a) social factors; (b) student factors; and, (c) school factors.

Social factors. The social factors that have been found to be associated with dropout behavior include poverty, coming from a single parent family, parents with a low level of education, and belonging to a language minority group (Ekstrom, Goertz, Pollack, & Rock, 1986; Fine, 1986; Johnston, Markle, & Harshbarger, 1986; McDill, Natriello, & Pallas, 1985; Morris et al., 1991; Rumberger, 1983; Steinberg, Blinde, & Chan, 1984; Wehlage & Rutter, 1986; Wehlage et al., 1989). Although it is not clear why individuals with the above characteristics are more likely to drop out of school, it has been hypothesized that the most likely contributing factors include: (a) fewer educational materials found in the home; (b) financial demands requiring older children to hold jobs or staying home to care for younger siblings; (c) parents who have had bad experiences in school and who do not value schooling; (d) parents with lower educational levels providing less assistance with homework; and, (e) language minority students who feel a stronger impact of the above factors because they cannot speak the dominant language (English or French in Canada), which is the instructional language of schools (Ekstrom et al., 1986; Fine, 1986; Rumberger, 1983; Steinberg et al., 1984).

Student factors. Johnston et al. (1986) reported that students who dropped out generally have lower occupational goals for themselves than students who have completed high school. Fine (1986) reported that students who dropped out are generally less motivated to perform in school and feel that although completing high school is important to finding a future job, school did not have anything important to offer to them. Catterall (1987) reported that dropout rates were much higher for those students who had expressed any doubts about finishing school. Furthermore, Johnston et al. (1986), Ekstrom et al. (1986), Wehlage and Rutter (1986), and Wehlage et al. (1989) reported that dropouts take part in fewer after school activities than their counterparts who obtain a high school diploma. This is partly a result of dropouts isolating themselves long before they finally leave school, and partly because a dropout is more likely to have an after school job where he or she is employed for more than 15 hours per week (Fine, 1986; McDill et al., 1985; Rumberger, 1987; Wehlage & Rutter, 1986).

Johnston et al. (1986), Strother (1986) and Kenney (1987) identified pregnancy or child related responsibilities as being one reason why female students drop out of school. However, pregnancy has been identified as a symptom rather than a cause of female dropout behavior. Girls who are performing poorly in school and who lack the basic skills associated with school success were found to be two and a half times more likely to become pregnant than females who were successful at school (Kenney, 1987).

School factors. One of the best predictors of school dropout behavior has been reported to be academic performance. Students with a poor academic history are much more likely to leave school prematurely than are students who are performing at or near grade level. Furthermore, students who dropped out of school were reported to have had poorer work habits and lower locus of control than graduates (Ekstrom et al., 1986; Rumberger, 1983; Wehlage & Rutter, 1986).

Hammack (1986), Lee (1986), Schulz, Toles and Rice (1987) and Shepard and Smith (1989) all reported that students who were overage for their grade

level (i.e., failed and repeated a grade), were more likely to leave school than were their peers who were at the appropriate age/grade level. School attendance has been shown to be another strong predictor of school dropout behavior. Students who eventually dropped out of school have been found to have had sporadic attendance long before they made the final decision to leave school (Broikowske, 1987; McDill et al., 1985; Wehlage & Rutter, 1986; Wehlage et al., 1989).

Thus far this review deals only with factors that are directly related to the student. However, additional factors relate directly to the educational systems where the students attended before dropping out. Researchers such as Fine (1986), Frymier and Gansmeder (1989), Johnston et al. (1986), Wehlage and Rutter (1986), and Wehlage et al. (1989) have reported that students drop out because schools do not offer them the programs nor the caring attention they need if they are to stay in school. School leavers reported that teachers were not really interested in them, nor did teachers understand their problems. For these students school was a place of failure that did not provide the skills relevant for their survival.

Present Research

The research carried out thus far has concentrated on developing profiles of the students who dropped out of school. This type of research provided much needed information as to the general characteristics of those male and female students who leave school early. However, none of the studies that we are aware of has attempted to discover why some students with all the identified general characteristics of a dropout actually drop out of school, whereas others graduate. Therefore, the purpose of this research was to compare dropouts and graduating students who have in common the same general characteristics of a dropout. Specifically, graduates and dropouts were compared regarding their perceptions toward teacher supports, family supports, their academic achievement, involvement in extracurricular activities, study habits, school curriculum, school climate, and overall satisfaction with their educational experience.

Method

Population

Three Winnipeg inner city high schools took take part in this study. These schools were selected because they represented a complete range of educational programing for inner city high school students. School A offered mainly a vocational program, School B offered a mixture of vocational and academic programs, and School C offered mainly an academic program.

School A. This school's enrollment was 275 students at the time the study was carried out and the population was made up of approximately 90% Aboriginal students. The median yearly family income was between $14,000 and $20,000. Sixty two percent of the students were from single parent families. Twenty percent spoke English as a second language. The unemployment rate for the parents of these students was 84% (Winnipeg School Division No. 1, 1990).

School B. Enrollment for this school at the time of the study was 848 students. The population was made up of approximately 60% Aboriginal, 10%

visible minorities (e.g., Portuguese, Filipino, Vietnamese), with the remainder white. The median income for the families of these students was between $14,000 and $20,000. Fifty seven percent of the students were from single parent families, with an unemployment rate of approximately 65%. With regard to the language spoken at home, approximately 11% of the families spoke English as a second language (Winnipeg School Division No. 1, 1990).

School C. The total enrollment for School C at the time of the study was 667 students. School C had the highest income range of the three, varying between approximately $20,001 and $30,000. This school had the lowest unemployment rate at approximately 25%. However, the population at this school had the largest ethnic mix including families of Aboriginal, Portuguese, Filipino, Vietnamese, Italian, Laotian, Chinese, and Polish descent. It also had the highest percentage of English as a second language students (29.1%). Approximately 29% of the students were from single parent families (Winnipeg School Division No. 1, 1990).

Sample recruitment: Graduates. The teaching staff at each of the three schools were asked to identify the students who were to graduate in June 1991, and all potential graduates were approached to take part in the study. For students under 18 years of age, signed parent permission letters were obtained before they were asked to complete the survey. The numbers of graduating students who took part from each of the three schools included seven students from School A, 40 students from School B, and 45 students from School C.

Sample recruitment: Dropouts. For the present study, a dropout was defined as any individual who had left school within the last three years and who would have graduated in June 1991 if they had not left school. However, the identification of the dropout sample did not proceed as expected, because locating dropouts who were willing to participate in the study proved to be difficult. It was hoped that a snowballing procedure using the in-school sample to identify friends who had dropped out would provide the required sample, but this procedure produced only three individuals. The dropouts identified by this method were not willing to complete the survey. Therefore, the majority of the dropouts included were recruited from four local programs designed to provide remedial academic assistance, life skill training, and career exposure to high school dropouts. The dropouts enrolled in these four programs had all left one of the three schools identified above. A total of 35 dropouts who matched the operational definition of a dropout agreed to participate.

Survey Instrument

This study used a 128-item survey that took approximately 55 minutes to complete. The following content areas were included: (a) personal data, which included age, gender, marital status, parental status, employment information, and work habits as a student; (b) school history and attitude toward school; (c) family background data, which included employment status, language spoken at home, parental educational attainment, and socioeconomic status; (d) out of school activities; (e) availability and use of family support structures such as space for doing homework, encouragement for completing high school, and so forth; and, (f) availability and use of school supports such

as counselors, support programs, peer groups, extra assistance from teachers, and so forth.

Some of the items were developed by the authors, whereas other items were taken with permission from Goldfarb's Questionnaire to Graduates (Karp, 1988) and from the Student Assessment Survey (Burstein & Sirotnik, 1984).

This instrument was reviewed by teachers from the three schools and was piloted with four students. Based on the feedback obtained from the teachers and the pilot students, a number of minor changes in question wording were made (a copy of the survey can be obtained from the first author).

Procedure

The study was carried out in May 1990. Parent permission letters were sent home with students who had been identified by their teachers as being most likely to graduate and who were under the age of 18 . For those students over the age of 18, the study was explained to them by the researchers and they were asked to sign a letter stating they agreed to take part in the study. The same procedure was used for the dropouts in the four community programs.

The students still attending school completed the survey during their class time. At School A, the surveys were administered by the researcher who read the questions to them, whereas at Schools B and C the teachers read the questions to the students. For the dropouts, the surveys were administered by the researcher who read the questions to them. The researchers trained the teachers in the administration of the survey in order to ensure that the survey was administered in a standardized manner to all individuals. All data were collected over two weeks.

Results

The data included completed or partly completed surveys from a total of 92 (46 male and 46 female) graduating students and 38 (27 male and 11 female) dropouts. The average age of the graduating students was 18.5, whereas for the dropouts it was 19.5. Of the 92 graduates, 46 (50.0%) spoke only English, whereas 14 (15.2%) also spoke an Aboriginal language, eight (8.7%) Polish, eight (8.7%) French, two (2.2%) Chinese, two (2.2%) Vietnamese, two (2.2%) Ukrainian, two (2.2%) Laotian, two (2.2%) Italian, two (2.2%) Filipino, one (1.1%) Hebrew, one (1.1%) Spanish, one (1.1%) Yugoslavian, and one (1.1%) Hungarian. Of the 38 dropouts, 18 (47.4%) spoke only English whereas eight (21.1%) also spoke an Aboriginal language, five (13.2%) Chinese, two (5.3%) Vietnamese, and three (7.9%) Cantonese.

Because of the high number of chi-square and t-tests carried out, alpha was set at the conservative .01 level in order to reduce the probability of making a Type I error (Hays, 1988).

Social Factors

Employment status. The results indicated that of the individuals who reported the employment status of their parents, 48.9% of the graduates and 23.7% of the dropouts reported that their mothers worked full time. Furthermore, 29.0% of the dropouts and 7.6% of the graduates reported that their mothers collected unemployment or welfare benefits. With regard to their

fathers, 58.7% of the graduates reported them as working full time, compared with 18.4% of fathers of dropouts.

A total of nine (23.7%) dropouts did not know whether their mother was working and 16 (42.1%) did not know if their father was working. For graduates, six (6.5%) did not know their mother's working status, 20 (21.7%) did not know their father's working status.

A total of 54.4% of the 92 graduates reported that they were living in a home owned by their parent(s), whereas 13.2% of the 38 dropouts reported living in a home that was owned by their parent(s).

Family structure. In comparing family structures, it was found that 57.6% of graduates reported living in two parent families whereas 21.1% of the dropouts reported living in two parent families. Of the graduates, 20.7% reported living with a guardian/foster parent, whereas 39.5% of the dropouts were living with a guardian/foster parent.

Language. Approximately equal percentages of dropouts (47.0%) and graduates (50.0%) reported that they could understand at least one language other than English. However, 36.8% of the dropouts and 26.0% of the graduates reported that they lived in homes where a language other than English was predominantly spoken.

Parents' educational level. Of the 92 graduates, 40.2% did not know their father's highest educational attainment, whereas 27.2% did not know their mother's highest educational attainment. For the 38 dropouts who took part in this study, 52.6% did not know their father's educational attainment, whereas 44.7% did not know their mother's educational attainment.

It was found that 30.5% of the fathers of graduates had obtained a grade 12 or higher education level, whereas 13.0% of the fathers of dropouts had done so. Furthermore, 39.2% of the graduates mothers had obtained a grade 12 or higher education level, whereas 15.8% of the mothers of dropouts had done so.

Student Factors

Marital status. A total of 14.1% of the graduates and 10.5% of the dropouts reported being married. Three graduates (3.3%) and three dropouts (7.9%) reported having children with an additional four graduates (4.4%) and one dropout (2.6%) pregnant at the time they took part in this study.

Student employment. Of the 38 dropouts, 46.7% reported that they had worked part time during their last year in school and an additional 21.1% reported having a full-time job just prior to dropping out of school. On the other hand, of the 92 graduates, 21.1% reported working at a part-time job and an additional 3.3% reported working full time.

Self-perceptions. A series of 12 questions asked graduates and dropouts to rate themselves as a student (see Table 1). Based on an analysis of that data, it was found that dropouts were more likely than graduates to rate themselves as being less smart (Chi-square(1)=11.19, p=.001), less ambitious (Chi-square(1)=15.51, p=.001), less helpful to other students (Chi-square(1)=7.4, p=.001), less responsible (Chi-square(1)=15.94, p=.001) and less competitive (Chi-square(1)=6.10, p=.01). Although not reaching statistical significance, it was also found that dropouts were somewhat less self-confident (Chi-

Table 1
Self Rating as a Student

Criteria	Not at all	Only a little bit	Quite a bit	Very much
Hardworking				
Graduates	3.3%	18.5%	51.1%	26.1%
Dropouts	2.6%	31.6%	57.9%	7.9%
Well organized				
Graduates	1.1%	27.2%	57.9%	14.1%
Dropouts	2.6%	31.6%	52.6%	13.2%
Smart				
Graduates	1.1%	13.0%	69.6%	12.0%
Dropouts	7.9%	34.2%	50.0%	7.9%
Friendly				
Graduates	0.0%	7.6%	29.4%	63.0%
Dropouts	0.0%	2.6%	50.0%	60.5%
Self-confident				
Graduates	2.2%	23.9%	48.9%	25.0%
Dropouts	7.9%	34.2%	47.4%	10.5%
Punctual				
Graduates	6.5%	21.7%	44.6%	26.1%
Dropouts	2.6%	31.6%	42.1%	23.7%
Shy				
Graduates	20.7%	41.3%	27.2%	9.8%
Dropouts	10.5%	65.8%	10.5%	13.6%
Ambitious				
Graduates	2.2%	14.1%	48.2%	33.1%
Dropouts	5.3%	44.7%	29.0%	21.1%
Helpful				
Graduates	0.0%	9.8%	43.5%	46.7%
Dropouts	2.6%	26.1%	50.0%	21.1%
Cooperative				
Graduates	0.0%	13.0%	38.0%	46.7%
Dropouts	0.0%	21.1%	42.1%	36.8%
Responsible				
Graduates	1.1%	4.4%	35.9%	57.6%
Dropouts	0.0%	31.6%	42.1%	26.3%
Competitive				
Graduates	9.8%	19.6%	41.3%	28.3%
Dropouts	13.2%	39.5%	31.6%	15.8%

square(1)=3.60, p=.05), and less shy (Chi-square(1)=2.26, p=.13) than were graduates (see Table 1).

Parental expectations and support. A total of 36 questions asked students to rate their parents' support of their educational efforts. Half concerned their perceptions of their father's support, whereas half were concerned with their perceptions of their mother's support. These items dealt with issues such as assistance and encouragement to complete homework, expectation of the effort put forth at school, hopes for the student after high school, encourage-

ment to attend college or university, and expectations to complete high school. To analyze these data, a composite score was developed by adding up the scores for each of the 36 questions. This procedure was used based on Whitehead's (1985) review that concluded that composite scoring was an acceptable method. The results of this procedure produced a mean score for graduates of 84.7 (SD=24.1) and a mean score for dropouts of 74.1 (SD=24.4) out of a total possible score of 144. Using an independence t-test with alpha set at .01, a significant difference between dropouts and graduates was found $t(128)=-2.25$, $p=.01$). Furthermore, when asked where they studied at home, 17.4% of the graduates and 34.2% of the dropouts stated that they "do not study at home." Therefore, these results suggest that graduates perceived that their parents were more supportive of them in their school activities than did dropouts.

School Factors

School attendance. From Table 2 it can be seen that although 42.4% of the 92 graduates had missed fewer than 10 days of school during their last year, 42.1% of the dropouts reported having missed more than 40 days. Furthermore, analysis of the data indicated that the reason most often given by 34.8% of the graduates for missing school was "illness," whereas for 47.4% of the dropouts the most often reported reason was "skipping" school. A total of 22.8% of the graduates gave "skipping" school as their reason for missing school, whereas 29.0% of the dropouts reported that "having a personal problem" was the reason behind most of their absences from school.

Grades repeated. When asked about the number of grades, if any, that they had repeated, 55.4% of the graduates reported that they had not repeated a grade whereas 36.8% of the school leavers reported that they had never repeated a grade. Surprisingly, the same percentage of dropouts and graduates (29%) reported that they had repeated one grade. However, 15.8% of the dropouts had repeated two grades and 15.8% had repeated more than two grades, whereas 5.4% of the graduates had repeated two grades and 6.5% had repeated more than two grades. The differences between graduates and dropouts regarding grades repeated were not found to be statistically different (Chi-square(1)=3.95, $p=.05$).

Courses offered. In order to compare graduate and dropout perceptions concerning the courses offered to them, participants were asked to rate those

Table 2
Attendance During Last Year of School

Attendance	*Graduates*	*Dropouts*
Missed less than 10 days	42.4%	13.2%
Missed between 10-20 days	27.2%	10.5%
Missed between 21-30 days	12.0%	21.0%
Missed between 31-40 days	13.0%	13.2%
Missed more than 41 days	5.4%	42.1%
Totals	100.0%	100.0%

offered to them, or that they had taken. A total of eight questions made up this section of the questionnaire. Question items included, for example, (a) were the courses offered the ones they wanted to take? (b) did the courses teach them what they needed to know? (c) was the difficulty level of the courses appropriate for them? and (d) was resource help available if they needed it? The participants were asked to rate the items on a four point scale from (1) strongly disagree to (4) strongly agree. To analyze these data, a composite score was developed by adding the scores for each of the eight questions. The results of this procedure produced a mean score for graduates of 25.0 (SD=3.2) and a mean score for dropouts of 23.0 (SD=4.0) out of a total possible score of 32. Using an independence t-test with alpha set for .01, a significant difference between dropouts and graduates was found ($t(128)=-3.06$, $p=.005$). Therefore, graduates perceived that the courses offered to them were closer to what they wanted to take than did dropouts.

Teachers as caring. Seven items were utilized to assess what difference, if any, there was between the graduates' and dropouts' perceptions of their teachers. These seven items asked participants to rate teachers on: friendliness, helpfulness, expectations, desire to communicate with them, whether they felt teachers cared about them, and whether they felt teachers did a good job. Compiling composite scores revealed that, out of a total possible score of 28, the mean for the graduates was 22.8 (SD=3.8) whereas the mean for the dropouts was 19.3 (SD=5.8). This difference was found to be statistically significant ($t(128)=-4.1$, $p=.005$). Therefore, graduates were found more likely to perceive their teachers as caring.

Student activities. Two sets of questions asked students about their extracurricular activities. The first set asked participants to indicate in which school activities, if any, they had participated. Table 3 presents the results of those questions. Although it can be seen that some differences do exist between graduates and dropouts, overall these differences were not found to be statistically significant (chi-square(1)= 0.324, $p=.32$). Further analysis of the data indicated that graduates participated in a mean of 2.5 activities, whereas dropouts participated in a mean of 2.3 activities, suggesting that graduates and dropouts participated in the same amount of extracurricular activities.

The second set of questions asked participants to indicate their involvement in a number of community activities. These questions, along with the results, are presented in Table 4. From that table it can be seen that whereas the graduates were more likely to be involved with church (chi-square(1)=8.63, $p<.001$) and volunteer work (chi-square(1)=10.33, $p<.001$), dropouts were more likely to be involved with drug/alcohol use (chi-square(1)=12.86, $p<.001$) and crime (chi-square(1)=10.84, $p<.001$).

Overall satisfaction with their school experience. A total of 12 questions asked graduates and dropouts to rate their overall satisfaction with their school experience. These items asked participants to rate their satisfaction with: the quality of the teachers, types of courses offered, opportunity to learn a career, number of students in their classes, school rules, efforts to encourage completion of school, and so forth. As with previous groupings, a composite score was obtained by adding the 12 items together. Out of a total possible score of 48, the mean score for graduates was 38.0 (SD=5.4) whereas the mean score for

Table 3
Extracurricular Activities Participated in

Activity	Participated in No	Yes
Sports teams		
Graduates	54.4%	45.7%
Dropouts	39.5%	60.5%
Cheerleading		
Graduates	94.6%	5.4%
Dropouts	97.4%	2.6%
Debating		
Graduates	83.7%	16.3%
Dropouts	92.1%	7.9%
Peer tutor		
Graduates	90.2%	9.8%
Dropouts	84.2%	15.8%
Student council		
Graduates	72.8%	27.2%
Dropouts	89.5%	10.5%
School newspaper		
Graduates	90.2%	9.8%
Dropouts	92.1%	7.9%
Band		
Graduates	90.2%	9.8%
Dropouts	81.6%	18.4%
TV productions		
Graduates	90.2%	9.8%
Dropouts	86.8%	13.2%
Drama		
Graduates	73.9%	26.1%
Dropouts	84.2%	15.8%
Honor assembly		
Graduates	87.0%	13.0%
Dropouts	81.6%	18.4%
Hobby clubs		
Graduates	73.9%	26.1%
Dropouts	79.0%	21.0%
Year book		
Graduates	80.3%	19.6%
Dropouts	68.4%	31.6%
Peer counselling		
Graduates	88.0%	12.0%
Dropouts	94.7%	5.3%
Other		
Graduates	75.0%	25.0%
Dropouts	94.7%	5.3%

Table 4
Community Involvement

Activity	Not involved	Rarely involved	Often involved	Very involved
Church				
Graduates	47.8%	26.0%	15.2%	8.8%
Dropouts	76.3%	21.0%	0.0%	0.0%
Community club				
Graduates	56.5%	21.7%	18.5%	2.2%
Dropouts	39.5%	23.7%	26.3%	10.5%
Cultural activities				
Graduates	53.3%	30.3%	12.0%	4.4%
Dropouts	55.3%	13.2%	15.8%	15.8%
Library				
Graduates	41.3%	34.8%	17.4%	4.4%
Dropouts	63.2%	15.8%	18.4%	2.6%
Volunteer work				
Graduates	33.7%	27.2%	28.3%	10.9%
Dropouts	57.9%	31.6%	10.5%	0.0%
Drug/alcohol use				
Graduates	47.8%	27.2%	19.6%	5.4%
dropouts	18.4%	23.7%	29.0%	29.0%
Crime				
Graduates	89.1%	6.5%	3.3%	0.0%
Dropouts	57.9%	21.1%	13.2%	7.9%

dropouts was 30.3 (SD=7.6). The results of an independent samples t-test found that graduates rated their school experience significantly higher than did dropouts ($t(128)=6.41$, $p=.001$).

Discussion

The purpose of this research was to provide a comparison of students from three inner city high schools who were about to graduate with their peers who had dropped out of school. However, few dropouts were willing to participate in the study, and this led to the procedure of using dropouts enrolled in community programs designed to upgrade their skill levels. Therefore, the results presented here must be viewed in light of the fact that the dropouts who participated in this study might represent a special group. That is, these dropouts, while not attending a formal school, were attempting to upgrade their skill levels. Different results might have been obtained from individuals who had dropped out of school and who were not in the process of upgrading their school related skills.

In addition, the results reported here were obtained from a relatively small number of individuals from the lower income areas of Winnipeg, and neither the schools nor the students were randomly selected. Therefore, (a) the results reported here cannot be generalized to other socioeconomic groups, and (b) while the schools were selected because they were representative of three different types of high school programming, care must be taken when gener-

alizing the results to individuals attending other inner city high schools. Finally, the relatively small sample size resulted in a lack of statistical power, which may have led to Type II errors.

The graduates and dropouts who took part in this study shared many of the same characteristics. That is, graduates as well as dropouts came from areas of an inner city that had high percentages of low income, single parent, and language minority families. Also, both groups were from inner city schools that had high numbers of students who had repeated a grade, high course failure rates, high dropout rates, and students who took part in few in-school extracurricular activities. However, while the results indicated that both groups had similar characteristics (i.e., same socioeconomic status, family backgrounds, etc.), the data mirrored those presented by other researchers that indicated that graduates had fewer of these problems than did peers who left school before obtaining their high school diploma.

Social factors. Although this study was carried out in the inner city section of Winnipeg, the results indicate that graduates were financially better off than were the individuals who had dropped out of school. That is, the parent(s) of graduates were less likely to be unemployed and collecting either social assistance or unemployment benefits, and were more likely to be living in a home that they owned. Furthermore, even though the schools contained high percentages of single parent families, graduates were more likely than dropouts to be living with both parents.

However, although the three schools had high proportions of English as a second language students, language differences were not found between graduates and dropouts. This finding was opposite to that of Steinberg et al. (1984) who reported that dropouts were more likely to be from language minority groups. On the other hand, the education level of parents did seem to be a factor. Graduates reported being more aware of their parents' educational attainments than were dropouts, and further that both their mothers and fathers were more likely to have a grade 12 or higher education levels. The fact that graduates were more aware of their parents' educational attainments suggested that education was more highly valued in those families. This possibility was supported by the fact that graduates reported that their parents were more supportive of their educational efforts and had higher expectations for them than was reported by dropouts.

Student factors. Graduates reported feeling more positive about themselves. They reported feeling smarter, more ambitious, and more helpful to other students than did the individuals who left school before obtaining their high school diploma.

Although graduates repeated fewer grades than did dropouts, this difference was not found to be statistically significant. However, dropouts did report that they skipped school more often than did graduates and also they were more likely to report that they had either a part-time or full-time job at the time they left school.

Although past research suggested that dropouts are less likely to be involved in extracurricular school activities, the results presented above indicate no difference between graduates and dropouts on this variable. However, a difference was found between these two groups for out of school activities.

That is, dropouts were more likely to have reported being involved with drug/alcohol and criminal activities, and less likely to be involved in volunteer and church activities. Furthermore, the fact that we were unable to effectively use the snowballing technique for locating dropouts suggested that graduates and dropouts do not keep in touch with one another once the dropout has left school. It might even be possible that these two groups did not associate with each other even while the dropouts were still attending school. These data suggest that dropouts tend to isolate themselves from their more successful peers (Wehlage, et al., 1989; Wehlage, 1990; Finn, 1989). Further research is needed in order to investigate the pre-dropout relationship between graduates and their peers who leave school early.

School factors. Overall, graduates reported being more satisfied with their school experiences and felt that teachers cared about them. Finally, graduates were also more likely to report that the courses the three schools offered to them were the courses that they wanted to take.

Conclusions

The results presented in this study provide further evidence that school dropout behavior is a complicated issue. Just because a student has repeated a grade, comes from a low income single parent family does not mean that he or she will leave school early. In fact, it was reported by Wehlage (1990) that in the United States "80% of students from single-parents homes do not drop out ... 66% of all dropouts are white; 42% attended suburban high schools; 86% had English as their home language; 71% had never repeated a grade" (p. 2). Those who took part in this study had many characteristics in common, yet some dropped out while others completed their schooling. Wehlage et al. (1989), Wehlage (1990), and Finn (1989) have hypothesized that, when taken together, the evidence concerning why students drop out of school strongly suggests that dropout behavior is a long-term process in which the student slowly becomes disenfranchised from school and the schooling process. Furthermore, these educators believe that the reliance on "'structural characteristics' such as SES or race" (Finn, 1989, p. 133) as the reasons behind dropout behavior is misleading. The model put forward by Finn (1989) called the Participation-Identification Model, suggested that school dropouts have not developed an identification or bonding to schools or the schooling process. Therefore, these students slowly withdraw both emotionally and physically from school. Furthermore, as these students slowly withdraw from school, they take little or no part in school activities either in or out of the classroom, thus increasing the probability that they will not learn to value schooling as an important aspect of their lives. If school does not hold any value, then the final decision to leave without the diploma is easier to make.

Although schools can do little, if anything, to improve the home lives of the students in their charge, they can attend to the fact that dropouts expressed dissatisfaction with their courses, especially with regard to the types of courses offered, course difficulty, and the perceived relevance of the courses to real life work experiences. Schools can develop and offer courses that challenge students and provide them with skills that they perceive as meaningful. Since dropouts also reported that they did not experience warmth and caring from

their teachers, and that they were unhappy generally about their school experiences, schools can attempt to remedy these problems by identifying and addressing the needs of their students.

> But unfortunately, many students report that typical school work is uninteresting. Many at-risk students claim that they do the same things over and over again. Remediation is the curriculum for these students and they see it as boring. Because there is an assumption that "basic skills" must be mastered first, school work fails to challenge students' problem solving skills and therefore lacks qualities that generate intrinsic interest. (Wehlage, 1990, p. 7)

However, as a review of the dropout literature will quickly confirm, there is no single solution to the problem of early school leaving behavior. There is not that magic prevention program which, when adopted by a school, will prevent additional students from dropping out of school. Unless schools are willing to provide a variety of prevention programs designed to reengage students into the learning process, these students will eventually drop out of school. A wide variety of options are needed because not all at-risk students have the same needs, nor do all at-risk students respond to a single set of conditions or curriculum.

> Given the diversity of problems at-risk students encounter, the schools in our study developed strategies in response to these impediments based on two insights: Programs should be matched to student needs and problems, and strategies should take advantage of student interests and strengths. (Wehlage et al., 1989, p. 224)

References

Beck, L., & Muia, J.A. (1980). A portrait of a tragedy: Research findings on the dropout. *High School Journal, 68*, 65-72.

Broikowske, D. (1987). *Truancy: A prelude to dropping out*. Bloomington, IN: National Educational Service.

Burstein, L., & Sirotnik, K.A. (1984). *The use of survey data in comprehensive information systems for local school improvements: Purposes, practices, problems*. UCLA: Center for the Study of Evaluation.

Catterall, J. (1987). An intensive group counseling dropout prevention intervention: Some cautions on isolating at-risk adolescents within high schools. *American Educational Research Journal, 24*, 521-540.

Ekstrom, R.B., Goertz, M.E., Pollack, M., & Rock, D.A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers College Record, 87*, 356-373.

Employment and Immigration Canada. (1990). *Youth: A national stay-in-school initiative*. Ottawa: Author.

Fine, M. (1986). Why urban adolescents drop into and out of public high school. *Teachers College Record, 87*, 393-409.

Finn, J.D. (1989). Withdrawing from school. *Review of Educational Research, 59*(2), 117-142.

Frymier, J., & Gansneder, B. (1989). The Phi Delta Kappa study of students at risk. *Phi Delta Kappan, 71*, 142-146.

Hammack, F.M. (1986). Large school systems' dropout reports: An analysis of definitions, procedures, and findings. *Teachers College Record, 87*, 324-341.

Hammons, R.A. (1988). *Relationships between interschool transfer and dropout rate*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hays, W.L. (1988). *Statistics* (4th ed.). Toronto, ON: Holt, Rinehart and Winston.

Johnston, J.H., Markle, G.C., & Harshbarger, M. (1986). What research says—About dropouts. *Middle School Journal, 17*, 8-11.

Karp, E. (1988). *The dropout phenomenon in Ontario secondary schools: A report to the Ontario study of the relevance of education and the issue of dropouts*. Toronto: Ontario Ministry of Education.

Kenney, A.M. (1987). Teen pregnancy: An issue for schools. *Phi Delta Kappan, 68*, 728-736.

Lee, J.D. (1986). Education of Native adolescents in inner-city schools. *Canadian Journal of Native Education, 13*(2), 22-26.

Mann, D. (1986a). Can we help dropouts: Thinking about the undoable. *Teachers College Record, 87*, 305-323.

McDill, E.L., Natriello, G., & Pallas, A.M. (1985). Raising standards and retaining students: The impact of the reform recommendations. *Review of Educational Research, 55*, 415-433.

Mincer, J. (1989). Human capital and the labour market: A review of the current research. *Educational Researcher, 18*(4), 27-34.

Morris, S., Pawlovich, W., & McCall, D. (1991). *Evaluating the effectiveness of school drop-out prevention strategies: Some suggestions for future research.* Ottawa: Canadian Education Association.

Rumberger, R.W. (1983). Dropping out of high school: The influence of race, sex, and family background. *American Educational Research Journal, 20*, 199-220.

Rumberger, R.W. (1987). High school dropouts: A review of the issues and evidence. *Review of Educational Research, 57*(2), 101-121.

Schulz, E.M., Toles, R., & Rice, W.K. (1987). The association of dropout rates with student attributes. In W.T. Denton (Ed.), *Dropouts, pushouts, and other casualties* (pp. 89-103). Bloomington, IN: Center on Evaluation, Development and Research: Phi Delta Kappa.

Shepard, L.A., & Smith, M.L. (1989). *Flunking grades: Research and policies on retention.* New York: Falmer.

Steinberg, L., Blinde, P.L., & Chan, K.S. (1984). Dropping out among language minority youth. *Review of Educational Research, 54*, 113-132.

Strother, D.B. (1986). Dropping out. *Phi Delta Kappan, 68*, 325-328.

Sullivan, M. (1988). *A comparative analysis of dropouts and non-dropouts in Ontario secondary schools.* Toronto: Ontario Ministry of Education.

Wehlage, G.G. (1990). The problem of identifying at-risk students. *River East School Division Journal, 2*(1), 1-10.

Wehlage, C.G., & Rutter, R.A. (1986). Dropping out: How much do schools contribute to the problem? *Teachers College Record, 87*, 374-392.

Wehlage, G.G., Rutter, R.A., Smith, G.A., Lesko, N., & Fernandez, R.R. (1989). *Reducing the risk: Schools as communities of support.* Philadelphia, PA: Falmer.

Whitehead, D.J. (1985). *The development and testing of an economics attitude scale.* Research papers in economics Education. London, UK: Institute of Education.

Winnipeg School Division No. 1. (1990). *1988-90 school description summary.* Winnipeg: Author.

For Contributors

Please prepare manuscripts in accordance with these guidelines.

Format

- Manuscript in triplicate.
- Typewritten or printed, double-spaced throughout.
- Abstract of approximately 100 words on a separate sheet.
- Manuscript not to exceed 6,500 words excluding graphics.
- Author's name and affiliation on a separate title page; title only on first page of manuscript to ensure anonymity in the review process.

Referencing

- Sources in parentheses after each reference (direct or otherwise), giving author's name (unless mentioned in text), year of publication, and page number(s).
- Direct quotes of 40 words or less in double quotation marks incorporated into text; those longer than 40 words in indented block format.
- All sources listed alphabetically at the end of the manuscript under the heading *References* using the APA style.
- Explanatory notes, numbered consecutively and marked in the text with superscript numerals, may appear before the References under the heading *Notes;* any citations in notes follow the same format as other references.

Graphics

- Each table and figure on a separate page at the end of the manuscript.
- Figures, charts, and diagrams in camera-ready form.
- Number tables in Arabic numerals and center over each heading, e.g., *Table 1.*
- Number figures in Arabic numerals before the caption at foot of each graph, e.g., *Figure 1: ...*
- Indicate placement of figures and tables in text, i.e., *Insert Table 3 here.*

Style

- *AJER*'s editorial style conforms closely to the *Publication Manual of the American Psychological Association.*
- Editorial changes may be made to manuscripts.
- For spelling, consult *Webster's New Collegiate Dictionary;* spelling in quoted material remains as in the original.

Diskettes

If manuscript is accepted for publication and has been prepared using a Macintosh, IBM, or IBM compatible microcomputer, please send diskette. Microsoft Word, MultiMate, WordPerfect, WordStar, Xerox Writer, XyWrite, and DisplayWrite 3 **only** are acceptable.
Important: Please indicate program used.

Submission to *AJER* implies the manuscript is not being considered for publication elsewhere.

AJER is indexed in the *Alberta Education Index, Canadian Education Index, Current Contents/Social and Behavioral Sciences, ERIC/Current Index to Journals in Education,* and *Social Science Citation Index;* appropriate articles are abstracted in *Cognitive Development Abstracts, Educational Administration Abstracts, Educational Technology Abstracts, Family Abstracts, Language Behavior Abstracts, Multicultural Education Abstracts, Psychological Abstracts, Research into Higher Education Abstracts, School Organization and Management Abstracts,* and *Sociology of Education Abstracts,* and reviewed by the *Journal for Research in Mathematics Education.*

If undelivered return to:

The Alberta Journal of Educational Research
4-116 Education North
University of Alberta
Edmonton, AB
T6G 2G5

Publications Mail Registration Number 1436